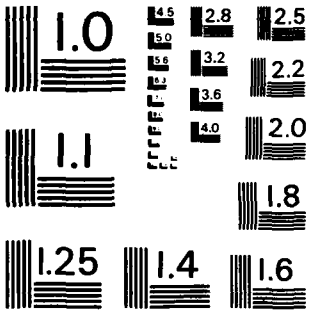
AD-A162 443 BOOTSTRAPPING COX'S REGRESSION MODEL(U) STANFORD UNIV CA LAB FOR COMPUTATIONAL STATISTICS N L HJORT NOV 85 LCS-TR-21 N00014-83-K-0472 1/1

UNCLASSIFIED F/G 12/1 NL

END



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

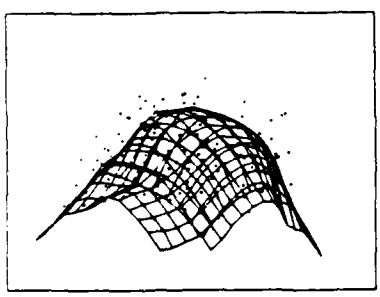
(6)

BOOTSTRAPPING COX'S REGRESSION MODEL

Nils Lid Hjort

Technical Report No. 21
November 1985

Laboratory for Computational Statistics



Department of Statistics
Stanford University

Bootstrapping Cox's regression model

by

Nils Lid Hjort

Norwegian Computing Centre

and

Department of Statistics, Stanford University

Abstract. Statistical inference in Cox's regression model is usually carried out using traditional (first order) large sample theory. In the spirit of earlier success stories one might try to bootstrap data in order to better assess the sampling variability of the Cox estimator. Such a bootstrap scheme is proposed in ~~the present~~ this paper. An asymptotic justification is given, showing that inference based on the bootstrap procedure is first order equivalent to the standard one. The problem of constructing more accurate moderate-sample confidence intervals is also addressed, employing second order fine-tuning of the bootstrap.

Key words: Bootstrap; Confidence interval; Cox model; Second order asymptotics.

Work supported by a National Science Foundation Grant MCS80-24649, Office of Naval Research contract N00014-83-K-0472.

1. Cox regression and a bootstrap method.

We consider a Cox regression model of the following form: $X_1^o,\ldots,X_n^o$ are independent lifetimes for n individuals. X_i^o has continuous intensity (or hazard rate)

$$\alpha_i(s) = \alpha(s)\, e^{\beta z_i} \tag{1.1}$$

where z_i is a covariate measurement for individual no. i, $\alpha(.)$ is left unspecified, and β is the parameter of primary interest. Observed is (X_i, δ_i), $i = 1,\ldots,n$, where

$$\begin{aligned} X_i &= \min\{X_i^o, c_i\}, \\ \delta_i &= I\{X_i^o \leq c_i\}. \end{aligned} \tag{1.2}$$

c_i is the "censoring time" for no. i.

We will assume, for simplicity and for ease of exposition, that z_i's and c_i's are non-random and that β (and z_i) is one-dimensional. Generalisations are possible in several directions, see Kalbfleisch and Prentice (1980), Andersen and Gill (1982), Prentice and Self (1982, 1983), Cox and Oakes (1984), Andersen and Borgan (1985). The important p-variate case is treated in Section 5.

Define

$$\begin{aligned} N_i(t) &= I\{X_i^o \leq t,\ X_i^o \leq c_i\}, \\ Y_i(t) &= I\{X_i^o \geq t,\ c_i \geq t\}. \end{aligned} \tag{1.3}$$

Cox' partial likelihood can be written

$$L(\beta) = \prod_{i=1}^{n} \prod_{s \geq 0} \left\{ \frac{Y_i(s)\exp(\beta z_i)}{\sum_{j=1}^{n} Y_j(s)\exp(\beta z_j)} \right\}^{dN_i(s)}, \tag{1.4}$$

cf. Gill (1984). Cox' estimator $\hat{\beta}$ is the β value maximising $L(\beta)$ or equivalently

$$\log L(\beta) = \sum_{i=1}^{n} \int_0^\infty [\beta z_i - \log \{ \sum_{j=1}^{n} Y_j(s)\, e^{\beta z_j} \}]\, dN_i(s).$$

This function may be seen to be concave so that $\hat{\beta}$ also may be defined as the solution to

$$U(\beta) = \partial \log L(\beta)/\partial\beta$$

$$= \sum_{i=1}^{n} \int_0^\infty \{z_i - \frac{S^{(1)}(s,\beta)}{S^{(0)}(s,\beta)}\} dN_i(s) = 0, \tag{1.5}$$

where

$$S^{(k)}(s,\beta) = \frac{1}{n} \sum_{j=1}^{n} (z_j)^k Y_j(s)\exp(\beta z_j), \quad k = 0, 1, 2. \tag{1.6}$$

We shall also need a nonparametric estimator of $A(t) = \int_0^t \alpha(s)ds$, the cumulative intensity for individuals with $z = 0$. The natural estimator is

$$\hat{A}(t) = \int_0^t \frac{d\bar{N}(s)}{\sum_{j=1}^{n} Y_j(s)\exp(\hat{\beta}z_j)}, \quad t \geq 0,$$

cf. Johansen (1983), Andersen and Borgan (1985), where $\bar{N}(.) = \sum_{j=1}^{n} N_j(.)$. This is a step function with jumps exactly at observed life-lengths, i.e. in

$$J = \{t: \Delta N_i(t) = 1 \text{ for some } i\}. \tag{1.7}$$

($\Delta B(t) = B\{t\} = B(t) - B(t-)$ for right continuous functions.) Since these jumps estimate conditional probabilities we modify the estimator above very slightly so as to get jumps

$$\Delta\hat{A}(s) = \min \{\frac{\Delta\bar{N}(s)/n}{S^{(0)}(s,\hat{\beta})}, 1\}, \quad s \in J. \tag{1.8}$$

Our main concern in this paper is assessing the variability of $\hat{\beta}$, by estimating its bias, its standard deviation, and by constructing confidence intervals for β based on $\hat{\beta}$. There exist answers to these questions based on large sample theory, reviewed in Section 2, but the approximations involved may be coarse. Recent success stories for the bootstrap, see for example Efron (1982a, 1985a, 1985b), Bickel and Freedman (1981), Singh (1981), Beran (1982), Freedman and Peters (1984), Abramovitch and Singh (1985), indicate that the sampling variability of $\hat{\beta}$ may be more accurately computed using bootstrap procedures.

Such a bootstrap method is now described. There are continuous c.d.f.s $F, F_1,\ldots,F_n$ having respectively $\alpha, \alpha_1,\ldots,\alpha_n$ of (1.1) as their intensities; in fact

$$\begin{aligned} F(t) &= 1 - \exp\{-A(t)\}, \\ F_i(t) &= 1 - \exp\{-A_i(t)\}, \end{aligned} \tag{1.9}$$

where $A, A_1,\ldots,A_n$ are the cumulative intensities. We could use

$$\begin{aligned} \tilde{F}(t) &= 1 - \exp\{-\hat{A}(t)\}, \\ \tilde{F}_i(t) &= 1 - \exp\{-\hat{A}(t)\exp(\beta z_i)\} \end{aligned}$$

as estimators, but prefer

$$\begin{aligned} \hat{F}(t) &= 1 - \Pi_{[0,t]}\{1 - \Delta\hat{A}(s)\}, \\ \hat{F}_i(t) &= 1 - \Pi_{[0,t]}\{1 - \Delta\hat{A}(s)\}^{\exp(\hat{\beta}z_i)}. \end{aligned} \tag{1.10}$$

$\hat{A}$ is a step function, corresponding to a probability distribution concentrated on the set J of (1.7), and this distribution is exactly $\hat{F}$ above, cf. Kalbfleisch and Prentice (1980, Ch. 2). Furthermore, the canonical analogue of (1.1) for discrete distributions is

$$1 - \Delta\hat{A}_i(s) = \{1 - \Delta\hat{A}(s)\}^{\exp(\hat{\beta}z_i)}, \tag{1.11}$$

cf. Kalbfleisch and Prentice (op. cit., Ch. 4). This leads to $\hat{F}_i$ above.

Now generate independent realisations

$$X_i^{o*} \sim \hat{F}_i, \quad i = 1,\ldots,n$$

and define

$$X_i^* = \min\{X_i^{o*}, c_i\}, \quad \delta_i^* = I\{X_i^{o*} \leq c_i\}. \tag{1.12}$$

$(X_1^*,\delta_1^*),\ldots,(X_n^*, \delta_n^*)$ is our bootstrap data. Calculate $\hat{\beta}^*$ as in (1.5), but based on the bootstrap sample.

$$\hat{G}(t) = \Pr_*\{\hat{\beta}^* \leq t\} \tag{1.13}$$

is the bootstrap distribution of $\hat{\beta}^*$, which in practice is obtained as

$$\hat{G}(t) \doteq \frac{1}{BOOT} \sum_{b=1}^{BOOT} I\{\hat{\beta}^{*b} \leq t\} \tag{1.14}$$

for a large number BOOT of independent realisations $\hat{\beta}^{*b}$ of $\hat{\beta}^*$. This requires of course computing power and time.

The bootstrap idea is that $\hat{G}$, the distribution of $\hat{\beta}^*$ given data, approximates G, the sampling distribution of $\hat{\beta}$, or more ambitiously that

$$Q(\hat{\beta}^*, \hat{\beta})|\text{data} \doteq Q(\hat{\beta}, \beta) \tag{1.15}$$

for any well-behaved function Q.

Summoning the necessary amount of courage to follow this principle, we can define

(A) The bias-corrected estimate $\bar{\beta}$: write $E(\hat{\beta} - \beta) = \text{bias}$. Since

$$(\beta^* - \hat{\beta})|\text{data} \doteq \hat{\beta} - \beta$$

the bootstrap estimate of bias is

$$\text{bias}_{BOOT} = E_*(\hat{\beta}^* - \hat{\beta}) \doteq \frac{1}{BOOT} \sum_{b=1}^{BOOT} (\hat{\beta}^{*b} - \hat{\beta}).$$

Hence

$$\bar{\beta} = \hat{\beta} + \text{bias}_{BOOT} \doteq \frac{1}{BOOT} \sum_{b=1}^{BOOT} \hat{\beta}^{*b}$$

should have less bias than $\hat{\beta}$.

(B) The bootstrap estimate of standard error is similarly

$$\hat{\sigma}_{BOOT} = \{\text{Var}_*(\hat{\beta}^*)\}^{\frac{1}{2}} \doteq \{\frac{1}{BOOT-1} \sum_{b=1}^{BOOT} (\hat{\beta}^{*b} - \bar{\beta})^2\}^{\frac{1}{2}}.$$

$\hat{\sigma}_{BOOT}$ intends to estimate $(\text{Var } \hat{\beta})^{\frac{1}{2}}$. One may also employ more robust versions of $\hat{\sigma}_{BOOT}$. That this sometimes is necessary is explained in Rey (1983) and Parr (1985), for example; see also a comment in Efron's reply to the discussants in Efron (1981b).

One may also estimate for example $E|\hat{\beta} - \beta|$ and $\{E(\hat{\beta} - \beta)^2\}^{\frac{1}{2}}$ in similar ways.

(C) The bootstrap percentile interval:

$$\hat{G}^{-1}(\alpha) \leq \beta \leq \hat{G}^{-1}(1-\alpha)$$

should have coverage probability close to $1-2\alpha$. This interval can further be corrected in various subtle ways; this is the topic of Section 4.

Section 2 reviews some known large sample theory for Cox regression. The same methods of proof, essentially elegant modern martingale techniques, are supplemented with brute force arguments in Section 3 in order to arrive at an asymptotic justification for the bootstrap procedure, for example

$$\sqrt{n}\ (\hat{\beta}^* - \beta) \,|\, \text{data} \approx \sqrt{n}\ (\hat{\beta} - \beta).$$

One may also prove the stronger result

$$\begin{pmatrix} \sqrt{n}\ (\hat{\beta}^* - \hat{\beta}) \\ \sqrt{n}\ \{\hat{A}^*(.) - \hat{A}(.)\} \end{pmatrix} \Big| \text{data} \approx \begin{pmatrix} \sqrt{n}\ (\hat{\beta} - \beta) \\ \sqrt{n}\ \{\hat{A}(.) - A(.)\} \end{pmatrix}.$$

The going gets tougher in Section 4 as we attempt to correct the percentile interval for bias and acceleration, which is the bootstrap way of doing second order asymptotics. Essentially Efron's (1985b) program is followed, but since the nuisance parameter $\alpha(.)$ is infinite-dimensional some intermediate analysis in an approximating finite-dimensional model is necessary.

An important ingredient in our construction of better intervals is the so called acceleration factor. It turns out that the calculation of this factor involves finding the skewness of a certain martingale, a technical problem solved in the Appendix, where also other mathematical necessities are dealt with. Some potentially interesting other uses of the skewness formula are pointed out in Section 5, along with some other remarks.

The results of our efforts is a confidence interval (4.18) that should be more accurate than the standard one, but which requires much more computation. The bootstrap sample size BOOT in (1.13), (1.14) should be at least 1000. Importance sampling and possibly smoothing tricks to estimate the necessary quantiles of $\hat{G}$ can perhaps be used to speed up convergence.

Remarks. (1) X_i^{o*} is simulated from $\hat{F}_i$, which has probability masses

$$\Delta\hat{F}_i(t) = \Pi_{[0,t)} \{1 - \Delta\hat{A}(s)\}^{\exp(\hat{\beta}z_i)} \Delta\hat{A}_i(t),$$
$$\Delta\hat{A}_i(t) = 1 - \{1 - \Delta\hat{A}(t)\}^{\exp(\hat{\beta}z_i)}, \tag{1.16}$$

for $t \in J$ of (1.7). We could also have used $\tilde{F}_i$ defined after (1.9), and the results would be the same in an asymptotic sense. However, looking at $\Delta\tilde{F}_i(t)$ reveals that the difference could be appreciable for small and moderate sample sizes, i.e. exactly in situations where the bootstrap is called for to fine-tune large sample approximations. Notice that $\hat{F}_i(t) > \tilde{F}_i(t)$ always.

The choice of $\hat{F}_i$, as opposed to $\tilde{F}_i$ or for example smoothed versions, is faithful to the original bootstrap spirit in that it reduces to the usual Kaplan-Meier estimate when $\hat{\beta} = 0$ (or when every $z_i = 0$) (and $\hat{A}$ is reduced to the usual Nelson-Aalen estimator), and is in this case also equal to the ordinary empirical distribution function when no censoring is present.

(2) The bootstrap scheme proposed above is particularly suited to the case where the censoring times c_i are known. (If there is no censoring at all then each $c_i = \infty$.) It utilises the information about c_i and z_i for individual no. i effectively. A typical example of this "fixed censorship" situation is one where individuals enter the experiment at different times, but where there is a fixed "data collection day", as was the case with the data Efron (1981a) considers.

In other situations the random censorship model could be more appropriate. Then $c_1, c_2, \ldots$ are considered as being independent of X_i^{o}'s and with a common c.d.f. R. Our bootstrap procedure should then be modified as follows: If $\delta_i = 0$, so that c_i is actually observed, use $c_i^* = c_i$. If $\delta_i = 1$, i.e. c_i is not observed, generate a likely outcome: $c_i^* \sim \hat{R}$, the Kaplan-Meier curve for R. Or, slightly more fancy: If $\delta_i = 1$ then one knows that the unobserved c_i is to the right of X_i, cf. (1.2), so one should perhaps generate c_i^* from $\hat{R}/\hat{R}(X_i,\infty)$. In any case one uses

$$X_i^* = \min\{X_i^{o*}, c_i^*\}, \quad \delta_i^* = I\{X_i^{o*} \leq c_i^*\}$$

instead of (1.12).

A related version could generate all $c_1^*,\ldots,c_n^*$ independently from $\hat{R}$; this would however not utilise the available information about the observed c_i's.

The resampling schemes discussed above simplify to those used by Efron (1981a) in the case of no covariates (or all covariates equal).

(3) If z_i is dichotomous (treatment/control, say) the bootstrap schemes simplify to close relatives of one used in Efron and Gong (1982, Section 7). More generally, if the z_i's take on K different values $z_{(1)},\ldots,z_{(K)}$ corresponding to K treatments or groups, then the bootstrap procedure amounts to generating X_i^{o*}'s in groups, say

$$X_{k,j}^{o*} \text{ i.i.d.} \sim \hat{F}_{(k)}(t) = 1 - \prod_{[0,t]} \{1 - \Delta\hat{A}(s)\}^{\exp(\hat{\beta}z_{(k)})}, \quad j = 1,\ldots,n_k$$

where n_k is the number of individuals having $z = z_{(k)}$, $k = 1,\ldots,K$.

Observe that the $X_{k,j}^{o*}$'s above are <u>not</u> resampled from the original observations, even when no censoring is present. Rather, they are generated from a better and model-based estimate $\hat{F}_{(k)}(.)$ of their true $F_{(k)}$ than is for example the usual empirical distribution for observations from group k.

(4) An important and related comment is that the Cox structure (1.1) is explicitly relied on in the sense that if the model is wrong, then $\hat{\beta}$ may display a different sampling variability. Thus the second order correct confidence intervals for β we construct in Section 4 are not necessarily even first order correct when the model is wrong (then the meaning of the "true" parameter β must be changed to the "least false" parameter; see Hjort (1985)). Similar remarks apply to the confidence intervals constructed under model assumptions in Efron (1985a, 1985b). More robust confidence intervals (for the least false parameter) would possibly be the result if one resampled the triplets (X_i, δ_i, z_i) directly, see the arguments in Efron (1982a, Section 5). This scheme is used in examples in Tibshirani (1984) and in Efron and Tibshirani (1985).

Resampling the triplets does not utilise the fine structure of the Cox model, and it is felt that the bootstrap scheme proposed here, based on the best available estimated model, is better suited to catching the finer aspects of the sampling variability of $\hat{\beta}$. It is still possible, however, that the two schemes are first order asymptotic equivalent, and that a second order correction to the simple method can match the second order correction we present in Section 4 for the model-utilising method.

It appears important to sort out the consequences (at least asymptotically) of using the simple method versus the model-based method, when the model is correct, and for specific departures from the model. This is not done here.

2. Asymptotic theory for Cox regression.

Introduce

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\exp(\beta_o z_i)\, dA(s), \quad i = 1,\ldots,n \tag{2.1}$$

where β_o denotes the true value of β. These are square integrable, orthogonal martingales with variance processes

$$(M_i, M_i)(t) = \int_0^t Y_i(s)\exp(\beta_o z_i)\, dA(s). \tag{2.2}$$

The family of σ-algebras implicitly referred to is

$$\mathcal{F}_t = \sigma\{N_i(s), Y_i(s) \; ; \; s \le t\}, \; t \ge 0.$$

Assume that the individuals are observed over a time interval $[0, T]$.

Then from (1.5)

$$U(\beta) = \sum_{i=1}^n \int_0^T \{z_i - \frac{S^{(1)}(s,\beta)}{S^{(0)}(s,\beta)}\}\, dM_i(s)$$

$$+ \sum_{i=1}^n \int_0^T \{z_i - \frac{S^{(1)}(s,\beta)}{S^{(0)}(s,\beta)}\}\, Y_i(s)\exp(\beta_o z_i)\, dA(s). \tag{2.3}$$

The second term vanishes when $\beta = \beta_o$, i.e. $U(\beta_o) = U(\beta_o, T)$ is a martingale, with

$$\frac{1}{n}(U(\beta_o), U(\beta_o))(T) = \frac{1}{n}\sum_{i=1}^n \int_0^T \{z_i - \frac{S^{(1)}(s,\beta_o)}{S^{(0)}(s,\beta_o)}\}^2\, d(M_i, M_i)(s)$$

$$= \int_0^T \{S^{(2)}(s,\beta_o) - \frac{S^{(1)}(s,\beta_o)^2}{S^{(0)}(s,\beta_o)}\}dA(s),$$

using (1.6). Assuming

$$S^{(k)}(s,\beta) \to s^{(k)}(s,\beta), \quad 0 \le s \le T, \; k = 0, 1, 2 \tag{2.4}$$

uniformly in s and in a neighbourhood of β_o, in probability, it follows that

$$n^{-\frac{1}{2}} U(\beta_o) \xrightarrow{d} N(0, \Sigma), \tag{2.5}$$

where

$$\Sigma = \sigma^2 = \int_0^T \{s^{(2)}(s,\beta_o) - \frac{s^{(1)}(s,\beta_o)^2}{s^{(0)}(s,\beta_o)}\}\, dA(s). \tag{2.6}$$

To employ a Taylor argument we need

$$I(\beta) = \partial^2 \log L(\beta)/\partial\beta^2 = \partial U(\beta)/\partial\beta$$

$$= -\sum_{i=1}^{n} \int_0^T \left[\frac{S^{(2)}(s,\beta)}{S^{(0)}(s,\beta)} - \left\{\frac{S^{(1)}(s,\beta)}{S^{(0)}(s,\beta)}\right\}^2\right] dN_i(s).$$

Using (2.1) once more we get

$$-\frac{1}{n} I(\beta) = \frac{1}{n} \sum_{i=1}^{n} \int_0^T \left[\frac{S^{(2)}(s,\beta)}{S^{(0)}(s,\beta)} - \left\{\frac{S^{(1)}(s,\beta)}{S^{(0)}(s,\beta)}\right\}^2\right] dM_i(s)$$

$$+ \int_0^T \left[\frac{S^{(2)}(s,\beta)}{S^{(0)}(s,\beta)} - \left\{\frac{S^{(1)}(s,\beta)}{S^{(0)}(s,\beta)}\right\}^2\right] S^{(0)}(s,\beta_o) dA(s).$$

It can be shown, using Lenglart's inequality, that if $\tilde{\beta} \xrightarrow{P} \beta_o$, then the first term goes to zero in probability, and

$$-\frac{1}{n} I(\tilde{\beta}) \xrightarrow{P} \Sigma. \tag{2.7}$$

This implies, using

$$0 = U(\hat{\beta}) = U(\beta_o) + I(\tilde{\beta})\ (\hat{\beta} - \beta_o)$$

for some $\tilde{\beta}$ between β_o and $\hat{\beta}$, that

$$\sqrt{n}\ (\hat{\beta} - \beta_o) = \{-\frac{1}{n} I(\tilde{\beta})\}^{-1}\ n^{-\frac{1}{2}}\ U(\beta_o)$$

$$\xrightarrow{d} \Sigma^{-1}\ N(0,\Sigma) = N(0,\ \Sigma^{-1}), \tag{2.8}$$

since $\hat{\beta}$ may be shown to consistent.

All this is in Andersen and Gill (1982) along with the necessary regularity conditions. That paper also establishes the consistency of the natural estimator

$$\hat{\Sigma} = \hat{\sigma}^2 = \int_0^T \{S^{(2)}(s,\hat{\beta}) - \frac{S^{(1)}(s,\hat{\beta})^2}{S^{(0)}(s,\hat{\beta})}\}\ d\hat{A}(s) \tag{2.9}$$

for Σ, and explores the simultaneous convergence in distribution of $\sqrt{n}\ (\hat{\beta} - \beta_o)$, $\sqrt{n}\ \{\hat{A}(.) - A(.)\}$.

3. Asymptotic justification for the bootstrap.

The aim of the present section is to show that inference based on the bootstrap method of Section 1 is asymptotically equivalent to the traditional inference that is now carried out each week on the basis of large sample results like those of Section 2. With such a result in the rear one could try to go further, exploring the possible superiority of the bootstrap to the traditional analysis in situations with smaller samples, by extensive simulation studies, as for example in Freedman and Peters (1984), or by theoretical investigations, for example along the lines of Beran (1982, 1984), Abramovitch and Singh (1985), Efron (1985a, 1985b). Some speculations of this sort are offered in Section 4.

With (1.15) representing the basic bootstrap idea we hope to compare the distribution of $\hat{\beta} - \beta_o$ with that of $\hat{\beta}^* - \hat{\beta}$ given data. We limit the discussion to sequences of outcomes where the Cox estimator $\hat{\beta}$ converges to the true β_o. This event, call it Ω_o, has probability 1 under the regularity conditions of Tsiatis (1981). It is not yet clear whether strong consistency of $\hat{\beta}$ can be established under weaker conditions of the type made by Andersen and Gill (1982); their martingale techniques yield only convergence in probability. We will henceforth assume sufficient regularity to ensure

$$\Pr(\Omega_o) = \Pr\{\hat{\beta} \to \beta_o\} = 1.$$

(Of course every random element, including covariate processes and censoring mechanisms, must then be defined on a proper common probability space. Without the strong consistency assumption the results below must be rephrased in a more cumbersome manner, and become of "in probability" type.)

We will avoid being too general here, and also avoid putting up all the needed regularity conditions; these are rather given implicitly by the arguments offered below. The i.i.d. like framework of Tsiatis (1981) will be sufficient;

it is suspected that also Andersen and Gill's conditions will make the arguments work.

The bootstrap sample is (X_i^*, δ_i^*), $i = 1,\ldots,n$ as in (1.12). Define

$$\begin{aligned} N_i^*(t) &= I\{X_i^* \le t,\ \delta_i^* = 1\}, \\ Y_i^*(t) &= I\{X_i^{o*} \ge t,\ c_i \ge t\}, \\ M_i^*(t) &= N_i^*(t) - \int_0^t Y_i^*(s)\, d\hat{A}_i(s) \\ &= N_i^*(t) - \Sigma_{[0,t]}\, Y_i^*(s)\Delta\hat{A}_i(s), \end{aligned} \tag{3.1}$$

where $\hat{A}_i$ is given by (1.8) and (1.11). The M_i^*'s become orthogonal martingales w.r.t.

$$F_t^* = \sigma\{N_i^*(s),\ Y_i^*(s)\ ;\ s \le t\},\ t \ge 0,$$

with variance processes

$$(M_i^*, M_i^*)(t) = \Sigma_{[0,t]}\, Y_i^*(s)\Delta\hat{A}_i(s)\{1 - \Delta\hat{A}_i(s)\}, \tag{3.2}$$

cf. Gill (1980) or Helland (1982). Furthermore $Y_i^*(.)$ is predictable (or previsible), i.e. $Y_i^*(t)$ is known at time t-.

$\hat{\beta}^*$ was defined as the solution to

$$U^*(\beta) = \sum_{i=1}^{n} \int_0^T \{z_i - \frac{S^{(1)*}(s,\beta)}{S^{(0)*}(s,\beta)}\}\, dN_i^*(s) = 0, \tag{3.3}$$

where

$$S^{(k)*}(s,\beta) = \frac{1}{n} \sum_{j=1}^{n} (z_j)^k\, Y_j^*(s)\exp(\beta z_j), \qquad k = 0,\ 1,\ 2. \tag{3.4}$$

We also need

$$\begin{aligned} I^*(\beta) &= \partial U^*(\beta)/\partial\beta \\ &= -\sum_{i=1}^{n} \int_0^T [\frac{S^{(2)*}(s,\beta)}{S^{(0)*}(s,\beta)} - \{\frac{S^{(1)*}(s,\beta)}{S^{(0)*}(s,\beta)}\}^2]\, dN_i^*(s). \end{aligned} \tag{3.5}$$

Proposition. With probability one,

$$\sqrt{n}\ (\hat{\beta}^* - \hat{\beta}) | \text{data} \to N(0,\ \Sigma^{-1}),$$

and both $\hat{\Sigma}^* = -\frac{1}{n} I^*(\hat{\beta}^*)$ and $\tilde{\Sigma}^* = -\frac{1}{n} I^*(\hat{\beta})$ converge to Σ in probability.

Indication of proof: The Taylor argument leading to (2.8) can be repeated to give

$$\sqrt{n}\ (\hat{\beta}^* - \hat{\beta}) = \{-\frac{1}{n} I^*(\tilde{\beta}^*)\}^{-1}\ n^{-\frac{1}{2}}\ U^*(\hat{\beta}), \tag{3.6}$$

where $\tilde{\beta}^*$ is between $\hat{\beta}$ and $\hat{\beta}^*$. We must prove that, for sequences in Ω_o, (i) $n^{-\frac{1}{2}}\ U^*(\hat{\beta}) \xrightarrow{d} N(0,\ \Sigma)$, (ii) $-\frac{1}{n} I^*(\tilde{\beta}) \xrightarrow{p} \Sigma$ whenever $\tilde{\beta} \xrightarrow{p} \beta_o$, and (iii) $\tilde{\beta}^* \xrightarrow{p} \beta_o$.

Basic to these results is the convergence

$$\Pr_*\{\sup_{0 \le s \le T,\ \beta \in B_o} |S^{(k)*}(s,\beta) - s^{(k)}(s,\beta)| \ge \varepsilon\ |\text{data}\} \to 0,\ \text{a.s., every } \varepsilon > 0, \tag{3.7}$$

i.e. $S^{(k)*}(s,\beta)$ ought to converge to the same $s^{(k)}(s,\beta)$ as did $S^{(k)}(s,\beta)$ in (2.4), uniformly in s and in a neighbourhood B_o of β_o, in probability, given data, for outcomes in Ω_o. This can be established using a (weak) law of large numbers for the space of right continuous functions with left hand limits on [0, T] to a separable Banach space, available from the proof of such a (strong) law given by Andersen and Gill (1982, Appendix II).

Start out writing

$$\begin{aligned} U^*(\beta) &= \sum_{i=1}^{n} \int_0^T \{z_i - \frac{S^{(1)*}(s,\beta)}{S^{(0)*}(s,\beta)}\}\ dM_i^*(s) \\ &\quad + \sum_{i=1}^{n} \int_0^T \{z_i - \frac{S^{(1)*}(s,\beta)}{S^{(0)*}(s,\beta)}\}\ Y_i^*(s) d\hat{A}_i(s) \\ &= U_1^*(\beta) + U_2^*(\beta) \end{aligned}$$

as in (2.3). Now $U_2^*(\beta)$ does not vanish for $\beta = \hat{\beta}$, but one may prove that

$$n^{-\frac{1}{2}}\ U_2^*(\hat{\beta}) \xrightarrow{p} 0,\ \text{a.s.},$$

by writing

$$\Delta\hat{A}_i(s) = 1 - \{1 - \Delta\hat{A}(s)\}^{\exp(\hat{\beta}z_i)}$$

$$= \Delta\hat{A}(s)\exp(\hat{\beta}z_i) - O_p\{\Delta\hat{A}(s)^2\}. \qquad (3.8)$$

(One could for simplicity assume the covariates to be bounded.) $\Delta\hat{A}(s)$ of (1.8) is of order $O_p(\frac{1}{n})$. Next look at $n^{-\frac{1}{2}} U_1^*(\hat{\beta})$. $U_1^*(\hat{\beta}) = U_1^*(\hat{\beta}, T)$ is a martingale (the processes $S^{(k)*}(s,\hat{\beta})$ are predictable in this conditional bootstrap framework), and

$$\frac{1}{n}(U_1^*(\hat{\beta}), U_1^*(\hat{\beta}))(T) = \frac{1}{n}\sum_{i=1}^{n}\int_0^T \{z_i - \frac{S^{(1)*}(s,\hat{\beta})}{S^{(0)*}(s,\hat{\beta})}\}^2 Y_i^*(s)\Delta\hat{A}_i(s)\{1 - \Delta\hat{A}_i(s)\}$$

$$\doteq \int_0^T \{S^{(2)*}(s,\hat{\beta}) - \frac{S^{(1)*}(s,\hat{\beta})^2}{S^{(0)*}(s,\hat{\beta})}\}\, d\hat{A}(s),$$

where $\doteq$ means ignoring $O_p\{\Delta\hat{A}(s)^2\}$ terms. That

$$n^{-\frac{1}{2}} U_1^*(\hat{\beta}) \xrightarrow{d} N(0, \Sigma), \quad \text{a.s.}$$

follows now by Rebolledo's central limit theorem for martingales, cf. Andersen and Gill (1982) or Helland (1982), using (3.7), $Pr(\Omega_o) = 1$, and a necessary lemma which states that

$$\int_0^T H_n(s)\, d\hat{A}(s) \xrightarrow{p} \int_0^T h(s)\, dA(s) \qquad (3.9)$$

if H_n is predictable and converges to h in probability. (3.9) is proved in the Appendix.

Now (i) stated in the beginning of the proof is demonstrated. (ii) and (iii) may be arrived at by similar efforts, following the route offered us by Andersen and Gill (1982) and repeatedly using arguments involving (3.8) and (3.9) when encountering new difficulties, and Lenglart's inequality. We will refrain from giving all the details here.

4. Confidence intervals.

The standard confidence interval for β is based on the large sample result $\sqrt{n}\ (\hat{\beta} - \beta)/\hat{\tau} \xrightarrow{d} N(0, 1)$, a consequence of (2.8) and (2.9), writing $\tau = 1/\sigma$, $\hat{\tau} = 1/\hat{\sigma}$. In fact,

$$\Pr_\beta\{\sqrt{n}\ (\hat{\beta} - \beta)/\hat{\tau} \leq t\} = \Phi(t) + O(n^{-\frac{1}{2}})$$

under mild extra conditions, for example boundedness of the covariates. Thus

$$\begin{aligned} \Pr_\beta\{\hat{\beta} - z^{(1-\alpha)}\hat{\tau}/\sqrt{n} \leq \beta\} &= 1 - \alpha + O(n^{-\frac{1}{2}}), \\ \Pr_\beta\{\hat{\beta} + z^{(1-\alpha)}\hat{\tau}/\sqrt{n} \leq \beta\} &= \alpha + O(n^{-\frac{1}{2}}), \end{aligned} \tag{4.1}$$

in particular the standard interval

$$\hat{\beta} - z^{(1-\alpha)}\hat{\tau}/\sqrt{n} \leq \beta \leq \hat{\beta} + z^{(1-\alpha)}\hat{\tau}/\sqrt{n} \tag{4.2}$$

has coverage probability $1 - 2\alpha + O(n^{-\frac{1}{2}})$. Here

$$z^{(\alpha)} = -z^{(1-\alpha)} = \Phi^{-1}(\alpha),\ z^{(1-\alpha)} = \Phi^{-1}(1-\alpha). \tag{4.3}$$

These confidence intervals (and similar ones for one out of several relative risk parameters $\beta_1, \ldots, \beta_p$) are widely used in biostatistics and the engineering sciences and can even make it to The New York Times.

The present section is concerned with the possibility of constructing confidence intervals for β with better moderate-sampling properties than the standard one.

Before we embark on that journey, let us briefly comment on the order of magnitude argument that led to (4.1). Results of A.I in the Appendix can be shown to imply that $\sqrt{n}\ (\hat{\sigma}^2 - \sigma^2)$ converges to some normal limit in distribution. It follows therefore from (2.8) that $\sqrt{n}\ (\hat{\beta} - \beta_o) = \{1/\sigma^2 + O_p(n^{-\frac{1}{2}})\}\ n^{-\frac{1}{2}}\ U(\beta_o)$, and the martingale $n^{-\frac{1}{2}}\ U(\beta_o)$ behaves in the correct way: It converges to a normal $(0, \sigma^2)$, and one may show using techniques of the Appendix that it has skewness $\gamma_{1,n}/\sqrt{n}$ where $\gamma_{1,n}$ tends to some γ_1, and kurtosis $\gamma_{2,n}/n$ where $\gamma_{2,n}$ tends to some γ_2. These facts surely indicate that the speed is the usual $\sqrt{n}$ towards normality, i.e. (4.1) is true. A more careful analysis, that perhaps

also could lead to a Berry-Esséen theorem for $\sqrt{n}\ (\hat\beta - \beta_o)/\hat\tau$, could start out rewriting $n^{-\frac12}\ U(\beta_o)$ as $A_n - n^{-\frac12}\ B_n$, where

$$A_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^T \{z_i - e(s,\beta_o)\}\ dM_i(s),$$

$$B_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^T Z_n(s)\ dM_i(s),$$

$$Z_n(s) = \sqrt{n}\ \{E(s,\beta_o) - e(s,\beta_o)\},$$

and where we for convenience have used $E(s,\beta) = S^{(1)}(s,\beta)/S^{(0)}(s,\beta)$ and $e(s,\beta)$ for its limit in probability $s^{(1)}(s,\beta)/s^{(0)}(s,\beta)$. The point of the rewriting is that A_n has independent summands, whereas $U(\beta_o)$ has dependent summands. If the covariates are bounded, then the n distributions governing the n martingales $M_1,\ldots,M_n$ are close to each other, and one can write down an Edgeworth-Cramér expansion or even a Berry-Esséen theorem for A_n. One may further show, using martingale techniques as in Section 2, that (A_n, B_n) tends in distribution to (A, B), say,where A and B are independent and Gaussian.

The brief discussion above can be made rigorous in the sense of securing (4.1), and can possibly also be used to establish (at least the existence of) an Edgeworth-Cramér expansion for $\sqrt{n}\ (\hat\beta - \beta_o)$.

4.1. Percentile and bias corrected bootstrap intervals.

The results of Section 3 imply $\sqrt{n}\ (\hat\beta^* - \hat\beta)/\hat\tau^* \xrightarrow{d} N(0, 1)$ a.s., where $(\hat\tau^*)^2 = \hat\Sigma^{*-1} = 1/\hat\sigma^{*2}$. This will imply that the bootstrap percentile interval

$$\hat G^{-1}(\alpha) \le \beta \le \hat G^{-1}(1-\alpha) \tag{4.4}$$

is first order asymptotically equivalent to the standard interval (4.2).

Let us in what follows suppose that sufficient regularity is in force to ensure

$$\Pr_*\{\sqrt{n}\ (\hat{\beta}^* - \hat{\beta})/\hat{\tau}^* \le t\} = \Phi(t) + O(n^{-1/2}) \text{ a.s.}, \tag{4.5}$$

writing $(\hat{\tau}^*)^2 = \hat{\Sigma}^{*-1}$, where the a.s. in the statement reminds us that the bootstrap probabilities $\Pr_*$ are conditioned on the random data. Uniform boundedness of the covariates is sufficient for this Berry-Esséen type statement to be true. (The arguments below are basically valid even without (4.5) and (4.1), but results must then be rephrased in a cumbersome manner less suited for exposition.) Then

$$\Pr_*\{\hat{\beta}^* \le \hat{\beta} + t\hat{\tau}^*/\sqrt{n}\} = \hat{G}(\hat{\beta} + t\hat{\tau}^*/\sqrt{n}) = \Phi(t) + O(n^{-1/2}) \text{ a.s.},$$

and one can show from this that

$$\hat{G}^{-1}(1 - \alpha) = \hat{\beta} + z^{(1-\alpha)}\hat{\tau}^*/\sqrt{n} + O(n^{-1}) \quad \text{a.s.},$$

$$\hat{G}^{-1}(\alpha) \quad = \hat{\beta} - z^{(1-\alpha)}\hat{\tau}^*/\sqrt{n} + O(n^{-1}) \quad \text{a.s.}$$

Hence (4.4) is first order equivalent to the standard interval (4.1), in particular

$$\Pr_\beta\{\hat{G}^{-1}(\alpha) \le \beta \le \hat{G}^{-1}(1 - \alpha)\} = 1 - 2\alpha + O(n^{-1/2}) \text{ a.s.} \tag{4.6}$$

There are ways to incorporate slight corrections to the percentile interval, the simplest of which is Efron's bias-corrected intervals, see Efron (1981b, 1982a, 1985a). Suppose that a smooth increasing transformation h exists with the property that

$$\sqrt{n}\ \{h(\hat{\beta}) - h(\beta)\} \dot{\sim} N(-b\lambda,\ \lambda^2), \tag{4.7}$$

$$\sqrt{n}\ \{h(\hat{\beta}^*) - h(\hat{\beta})\} \dot{\sim}_* N(-b\lambda,\ \lambda^2), \tag{4.8}$$

where $\dot{\sim}$ and $\dot{\sim}_*$ mean "approximately distributed as". b could be zero, in fact Sections 2 and 3 in a sense show that b will be zero, asymptotically, to the first order. We may think of b above as a second order fine-tuning of this, say $b = b_o/\sqrt{n}$.

The normal pivotal assumptions (4.7), (4.8) give a bias-corrected percentile interval, as in Efron (1982a,Ch. 10). The result is

$$\hat{G}^{-1}\{\Phi(z^{(\alpha)} + 2b)\} \le \beta \le \hat{G}^{-1}\{\Phi(z^{(1-\alpha)} + 2b)\} \tag{4.9}$$

where $b = \Phi^{-1}\{\hat{G}(\hat{\beta})\}$ compensates for the possible bias of the observed $\hat{\beta}$ as an estimate of β. If $\Pr_*\{\hat{\beta}^* \leq \beta\} \doteq \Pr\{\hat{\beta} \leq \beta\} = \frac{1}{2}$ then $b = 0$ and (4.9) reduces to the ordinary percentile interval.

It is not obvious that b can be taken to be the same in (4.7), (4.8); this is rather an optimistic (but educated) guess trusting the bootstrap approximation idea. There are also theoretical reasons for believing in a common $b = b_o/\sqrt{n}$, however. Further comments on this and the question of how precise the $\dot{\sim}$, $\dot{\sim}_*$ statements in (4.7), (4.8) must be are offered in the next subsection.

4.2. Second order correct intervals.

A competing interval

$$\hat{\beta}_{LOW} \leq \beta \leq \hat{\beta}_{UP}$$

to (4.2) and (4.4) would be more accurate if

$$\Pr_\beta\{\hat{\beta}_{LOW} \leq \beta\} = 1 - \alpha + O(n^{-1}),\ \Pr_\beta\{\hat{\beta}_{UP} \leq \beta\} = \alpha + O(n^{-1}). \tag{4.10}$$

Perhaps many intervals share this property; we are only interested in those that are "inferentially correct", which we take to mean "based on the Cox estimator $\hat{\beta}$". ($\hat{\beta}$ is an asymptotically optimal estimator according to Begun et al. (1983) (but shares this property also with for example Bayes estimators, see Hjort (1985).) We should therefore look for intervals of the type $\hat{\beta}_{LOW} = \hat{\beta} - z^{(1-\alpha)}\hat{\tau}/\sqrt{n} + A_n^{(\alpha)}/n$, $\hat{\beta}_{UP} = \hat{\beta} + z^{(1-\alpha)}\hat{\tau}/\sqrt{n} + A_n^{(1-\alpha)}/n$ with the A_n's chosen to make (4.10) true. Such intervals could be termed second order correct.

It is clear that second order asymptotics in some form or another must enter the discussion if such sharper intervals are to be constructed. The approach described now is due to Efron (1985b). The following may be read as essentially a review of his method, but with more attention paid to order of magnitude arguments and to the assumptions actually used to make the second order statements true.

The results of Section 2 imply via the delta method the more general first order asymptotic result

$$\Pr_\beta[\sqrt{n}\ \{g(\hat{\beta}) - g(\beta)\}/\lambda(\beta) \leq t] = \Phi(t) + O(n^{-\frac{1}{2}})$$

for each smooth transformation g, with $\lambda(\beta) = |g'(\beta)|\tau$. This is not sharp enough to get to (4.10). But assume that for some smooth increasing g and some cleverly chosen constants a_o, b_o

$$\sqrt{n}\,\{g(\hat{\beta}) - g(\beta)\}/\lambda(\beta) \doteq N(-b_o/\sqrt{n},\ 1), \qquad \lambda(\beta) = \lambda(\beta_o) + a_o\{g(\beta) - g(\beta_o)\}$$

in the sharper second order sense, i.e.

$$\Pr_\beta[\sqrt{n}\,\{g(\hat{\beta}) - g(\beta)\}/\lambda(\beta) \leq t] = \Phi(b_o/\sqrt{n} + t) + O(n^{-1}),$$

where $\lambda(\beta_o)$ is the standard deviation for the limiting distribution of $\sqrt{n}\,\{g(\hat{\beta}) - g(\beta)\}$ under some reference parameter value β_o. In order to improve chances of there being such a sharper g both a bias-correction in the form of $b_o/\sqrt{n}$ and an acceleration-correction entering the standard deviation have been allowed. The limiting standard deviation of $\sqrt{n}\,\{g(\hat{\beta}) - g(\beta)\}$ under β is of course not always of the form $\lambda(\beta_o) + a_o\{g(\beta) - g(\beta_o)\}$, we would in our situation expect it to depend also upon the unknown hazard function $\alpha(.)$ for example. However, the statement needs only be local in character, i.e. valid for β in a neighbourhood of the reference point β_o, and the dependence upon $\alpha(.)$ and possibly covariate values may be subsumed in the (approximate) constant a_o.

There is no harm in taking $g(\beta_o) = 0$ and $\lambda(\beta_o) = 1$ since this can be arranged by adjusting g accordingly. We therefore state the assumption above as

$$H_n(t) = \Pr_\beta\{\frac{\sqrt{n}\,(\hat{\gamma} - \gamma)}{1 + a_o\gamma} \leq t\} = \Phi(b_o/\sqrt{n} + t) + O(n^{-1}), \tag{4.11}$$

writing $\gamma = g(\beta)$, $\hat{\gamma} = g(\hat{\beta})$. (The O-term depends upon both β and t; the extent to which uniformity is needed is discussed later.) Notice the similarity to Edgeworth expansions; the r.h.s. may be written $\Phi(t) + \phi(t)\beta_o/\sqrt{n} + O(n^{-1})$.

Under assumption (4.11) a second order correct confidence interval for β, involving g, a_o, b_o, may be constructed. One has

$$\hat{\gamma} - \gamma \overset{d}{=} \frac{1}{\sqrt{n}} (1 + a_o\gamma)\ (Z - b_o/\sqrt{n})$$

where Z is very close to being standard normal,

$$\begin{aligned} \Pr_\beta\{Z - b_o/\sqrt{n} \le t\} &= \Phi(t + b_o/\sqrt{n}) + O(n^{-1}), \\ \Pr_\beta\{Z \le t\} &= \Phi(t) + O(n^{-1}). \end{aligned} \tag{4.12}$$

It follows that

$$1 + a_o\hat{\gamma} \overset{d}{=} (1 + a_o\gamma)\ \{1 + \frac{a_o}{\sqrt{n}} (Z - \frac{b_o}{\sqrt{n}})\}.$$

It should be safe to take logarithms since both γ and $(Z - b_o/\sqrt{n})/\sqrt{n}$ are close to zero; see Efron (1982b, Sections 4 and 8) for a more careful treatment, and the corresponding discussion in Efron (1985b). Hence

$$h(\hat{\gamma}) \overset{d}{=} h(\gamma) + W, \quad W = \log\ \{1 + \frac{a_o}{\sqrt{n}} (Z - \frac{b_o}{\sqrt{n}})\} \tag{4.13}$$

where $h(t) = \log\ (1 + a_o t)$. The natural interval on this $h\{g(\beta)\}$ scale becomes

$$h(\hat{\gamma}) - w^{(1-\alpha)} \le h(\gamma) \le h(\hat{\gamma}) - w^{(\alpha)},$$

where $\Pr_\beta\{W \le w^{(\alpha)}\} = \alpha$, $\Pr_\beta\{W \le w^{(1-\alpha)}\} = 1 - \alpha$. Some analysis shows that

$$w^{(\alpha)} = \log\ \{1 + \frac{a_o}{\sqrt{n}} (z^{(\alpha)} - \frac{b_o}{\sqrt{n}})\} + O(n^{-3/2}),$$

and similarly for $w^{(1-\alpha)}$. Using $h^{-1}(t) = (1/a_o)\ (e^t - 1)$ and some algebra

$$\begin{aligned} \tilde{\gamma}_{UP} &= h^{-1}\{h(\hat{\gamma}) - w^{(\alpha)}\} \\ &= \gamma_{UP} + O(n^{-3/2}) \end{aligned}$$

where

$$\gamma_{UP} = \hat{\gamma} + \frac{1}{\sqrt{n}} (1 + a_o\hat{\gamma})\ \frac{z^{(1-\alpha)} + b_o/\sqrt{n}}{1 - (a_o/\sqrt{n})\ \{z^{(1-\alpha)} + b_o/\sqrt{n}\}},$$

and similarly for $\tilde{\gamma}_{LOW} = \gamma_{LOW} + O(n^{-3/2})$.

Our "correct" interval is now

$$\beta_{LOW} \le \beta \le \beta_{UP} \; ; \qquad \beta_{LOW} = g^{-1}(\gamma_{LOW}),\ \beta_{UP} = g^{-1}(\gamma_{UP}). \tag{4.14}$$

One has by (4.13)

$$\begin{aligned}
Pr_\beta\{\beta_{UP} \le \beta\} &= Pr_\beta\{\gamma_{UP} \le \gamma\} \\
&= Pr_\beta\{h(\tilde{\gamma}_{UP} - O(n^{-3/2})) \le h(\gamma)\} \\
&= Pr_\beta\{h(\hat{\gamma}) - w^{(\alpha)} + O(n^{-3/2}) \le h(\gamma)\} \\
&= Pr_\beta\{W \le w^{(\alpha)} + O(n^{-3/2})\} \\
&= Pr_\beta[Z - b_o/\sqrt{n} \le \{\sqrt{n}/a_o\}\{\exp(w^{(\alpha)}) - 1\} + O(n^{-1})] \\
&= \Phi[b_o/\sqrt{n} + \{\sqrt{n}/a_o\}\{\exp(w^{(\alpha)}) - 1\} + O(n^{-1})] + O(n^{-1}) \\
&= \Phi\{z^{(\alpha)} + O(n^{-1})\} + O(n^{-1}) = \alpha + O(n^{-1}),
\end{aligned}$$

and similarly

$$Pr_\beta\{\beta_{LOW} \le \beta\} = 1 - \alpha + O(n^{-1}).$$

The interval (4.14) depends upon the unknown quantities g, a_o, b_o, and approximations to β_{LOW} and β_{UP} must be devised. (4.14) was derived under assumption (4.11); its bootstrap version is

$$H_n^*(t) = Pr_*\{\frac{\sqrt{n}\ (\hat{\gamma}^* - \hat{\gamma})}{1 + a_o\hat{\gamma}} \le t\} = \Phi(b_o/\sqrt{n} + t) + O(n^{-1}) \text{ a.s.}, \tag{4.15}$$

writing $\hat{\gamma}^* = g(\hat{\beta}^*)$, $\hat{\gamma} = g(\hat{\beta})$. Assuming (4.15) to be true and remembering (1.13),

$$\begin{aligned}
\hat{G}(\hat{\beta}) &= Pr_*\{\hat{\beta}^* \le \hat{\beta}\} \\
&= H_n^*(0) = \Phi(b_o/\sqrt{n}) + O(n^{-1}) \text{ a.s.}, \\
\hat{G}(\beta_{UP}) &= Pr_*\{g(\hat{\beta}^*) \le g(\beta_{UP})\} \\
&= Pr_*[\hat{\gamma}^* - \hat{\gamma} \le \frac{1}{\sqrt{n}}(1 + a_o\hat{\gamma})\ \frac{z^{(1-\alpha)} + b_o/\sqrt{n}}{1 - (a_o/\sqrt{n})\ \{z^{(1-\alpha)} + b_o/\sqrt{n}\}}] \\
&= H_n^*[\frac{z^{(1-\alpha)} + b_o/\sqrt{n}}{1 - (a_o/\sqrt{n})\ \{z^{(1-\alpha)} + b_o/\sqrt{n}\}}] \\
&= \Phi(z^{[1-\alpha]}) + O(n^{-1}) \text{ a.s.},
\end{aligned}$$

$$\hat{G}(\beta_{LOW}) = H_n^*\left[\frac{z^{(\alpha)} + b_o/\sqrt{n}}{1 - (a_o/\sqrt{n})\{z^{(\alpha)} + b_o/\sqrt{n}\}}\right]$$

$$= \Phi(z^{[\alpha]}) + O(n^{-1}) \text{ a.s.},$$

where

$$z^{[1-\alpha]} = \frac{b_o}{\sqrt{n}} + \frac{z^{(1-\alpha)} + b_o/\sqrt{n}}{1 - (a_o/\sqrt{n})\{z^{(1-\alpha)} + b_o/\sqrt{n}\}} \tag{4.16}$$

and analogously for $z^{[\alpha]}$. Hence

$$\left.\begin{aligned} b_o/\sqrt{n} &= \Phi^{-1}\{\hat{G}(\hat{\beta}) + O(n^{-1})\} = \Phi^{-1}\{\hat{G}(\hat{\beta})\} + O(n^{-1}) \text{ a.s.}, \\ \beta_{UP} &= \hat{G}^{-1}\{\Phi(z^{[1-\alpha]}) + O(n^{-1})\} = \hat{G}^{-1}\{\Phi(z^{[1-\alpha]})\} + O(n^{-3/2}) \text{ a.s.}, \\ \beta_{LOW} &= \hat{G}^{-1}\{\Phi(z^{[\alpha]}) + O(n^{-1})\} = \hat{G}^{-1}\{\Phi(z^{[\alpha]})\} + O(n^{-3/2}) \text{ a.s.} \end{aligned}\right\} \tag{4.17}$$

All this leads us to propose

$$\hat{\beta}_{LOW} = \hat{G}^{-1}\{\Phi(\hat{z}^{[\alpha]})\} \leq \beta \leq \hat{G}^{-1}\{\Phi(\hat{z}^{[1-\alpha]})\} = \hat{\beta}_{UP} \tag{4.18}$$

as the "final" interval, where

$$\hat{z}^{[1-\alpha]} = \frac{\hat{b}_o}{\sqrt{n}} + \frac{z^{(1-\alpha)} + \hat{b}_o/\sqrt{n}}{1 - (\hat{a}_o/\sqrt{n})\{z^{(1-\alpha)} + \hat{b}_o/\sqrt{n}\}} \tag{4.19}$$

and correspondingly for $\hat{z}^{[\alpha]}$, where $\hat{b}_o/\sqrt{n} = \Phi^{-1}\{\hat{G}(\hat{\beta})\}$, and where $\hat{a}_o/\sqrt{n}$ is an estimate of $a_o/\sqrt{n}$, a separate and difficult problem returned to in the following subsection.

The resulting interval (4.18) does <u>not</u> depend upon the transformation g or upon b_o, a_o, and is Efron's acceleration and bias corrected (ABC) interval. Observe that if n is large, then $\hat{z}^{[1-\alpha]}$ is close to $z^{(1-\alpha)}$ and $\hat{z}^{[\alpha]}$ is close to $z^{(\alpha)}$, i.e. (4.18) is not very different from the simple percentile interval (4.4), which in 4.1 was found to be first order equivalent to the standard interval (4.2). (4.18) purports to make the necessary next order corrections.

Each of the proposed estimates $\hat{a}_o$ of 4.3 has the property that $\hat{a}_o - a_o = O_p(n^{-\frac{1}{2}})$. Also, $\hat{b}_o/\sqrt{n} - b_o/\sqrt{n} = O(n^{-1})$ a.s. so that

$$\hat{z}^{[1-\alpha]} = z^{[1-\alpha]} + O_p(n^{-1}),$$

entailing

$$\left.\begin{aligned} \hat{\beta}_{UP} &= \hat{G}^{-1}\{\Phi(z^{[1-\alpha]}) + O_p(n^{-1})\} \\ &= \hat{G}^{-1}\{\Phi(z^{[1-\alpha]})\} + O_p(n^{-3/2}) \\ &= \beta_{UP} + O_p(n^{-3/2}), \\ \hat{\beta}_{LOW} &= \beta_{LOW} + O_p(n^{-3/2}). \end{aligned}\right\} \quad (4.20)$$

Suppose that the normalising transformation g is nearly perfect, i.e. the $O(n^{-1})$ term of (4.11) is practically zero. Statisticians have experienced wondrous g transformations like that in many situations, for example Fisher's $\tanh^{-1}$ transformation of the correlation coefficient and Wilson-Hilferty's cube root of the chi squared. Then the pivotal statement (4.13) is in force with a distribution for W that is really independent of the unknown parameters β and A, and the interval constructed from it, $h(\hat{\gamma}) - w^{(1-\alpha)} \leq h\{g(\beta)\} \leq h(\hat{\gamma}) - w^{(\alpha)}$ would be the <u>correct interval</u> in a strong sense, both inferentially and probabilistically, cf. the discussion following (4.10). What we have actually shown is that (4.18) has endpoints coming very close to the correct endpoints $g^{-1}[h^{-1}\{h(\hat{\gamma}) - w^{(1-\alpha)}\}]$, $g^{-1}[h^{-1}\{h(\hat{\gamma}) - w^{(\alpha)}\}]$, namely erring by just $O_p(n^{-3/2})$.

This seems to be Efron's motivation for and justification for the ABC interval (4.18). A perhaps separate issue is whether (4.18) is second order correct in the sense of the discussion following (4.10). We have come very close to establishing (4.10) too. Remarks (2) and (3) below explain why the key assumptions (4.11) and (4.15) can be trusted, and these assumptions were shown to imply that $\beta_{LOW} \leq \beta \leq \beta_{UP}$ had the prestigious $O(n^{-1})$ property. The

question is whether $\hat{\beta}_{UP} = \beta_{UP} + O_p(n^{-3/2})$ implies $Pr_\beta\{\hat{\beta}_{UP} \leq \beta\} = Pr_\beta\{\beta_{UP} \leq \beta\} + O(n^{-1})$. This certainly looks reasonable, and should be true under very mild conditions. A formal proof might perhaps need the existence of Edgeworth-Cramér expansions ror $\sqrt{n}\ (\hat{\beta}_{UP} - \beta)/\hat{\tau}$ and $\sqrt{n}\ (\beta_{UP} - \beta)/\hat{\tau}$, cf. the discussion in the beginning of Section 4.

Remarks. (1) From (4.16) we see that the confidence interval just as conveniently can be given in terms of $a = a_o/\sqrt{n}$ and $b = b_o/\sqrt{n}$ instead of a_o and b_o. We used a_o and b_o only for motivational purposes and for keeping track of the various oreders of magnitude involved in the discussion. Indeed a_o, b_o, and g are allowed to depend upon n too, but in a "stable" manner.

(2) The assumption (4.11) is not as restricitive at it may appear to be. It states that the distribution of $\hat{\beta}$, to a second order approximation, is a scaled normal translation family, in the language of Efron (1982b). It is an implicit result of Efron (1985b) and an explicit result of DiCiccio and Tibshirani (1985) that such a transformation always exists for one-parameter problems, and also, in appropriate senses, in multi-parameter and nonparametric models. The precise definitions behind these "appropriate senses" are (in their current formulation) evasive and circumventive in nature, however, and involve least favourable reductions to one-parameter situations, and are perhaps not entirely satisfactory. We shall indeed follow Efron (1985b) and DiCiccio and Tibshirani (1985) and employ one-parameter reduced models in the following subsection, where we struggle to find estimates for $a = a_o/\sqrt{n}$, and essentially use the arguments of 4.2 in the reduced model. Of course the Cox model fits neither of the categories mentioned above, because of its infinite-dimensional nuisance parameter. It can however be closely approximated with one with finitely many parameters, as explained in 4.3 below. On these grounds assumption (4.11) can be trusted.

(3) It is not at all obvious that (4.11) implies the other key assumption (4.15). H_n and H_n^* are close, and indeed the efforts of Section 3 establish $\sup_t |H_n(t) - H_n^*(t)| \to 0$ a.s. Needed now is a sharper statement, for example

$$\sqrt{n} \sup_t |H_n(t) - H_n^*(t)| \to 0 \text{ a.s.}$$

or ideally (pointwise) a.s. boundedness of $n|H_n(t) - H_n^*(t)|$.

Results of Bickel and Freedman (1980, 1981), Singh (1981), and Babu and Singh (1983) indicate that indeed $H_n^*(t) = H_n(t) + O(n^{-1})$ a.s. A lesson learned from these papers is that such results should not be taken for granted, however, and that small changes in the definition of statistics may result in drastic asymptotic differences. For example, it may be true that the distribution functions H_n^o and H_n^{o*} of the non-standardised variables $\sqrt{n}\,(\hat{\gamma} - \gamma)$ and $\sqrt{n}\,(\hat{\gamma}^* - \hat{\gamma})$, respectively, have $\sqrt{n} \sup_t |H_n^o(t) - H_n^{o*}(t)| = O(\log \log n) \to \infty$ a.s. even though $\sqrt{n} \sup_t |H_n(t) - H_n^*(t)| \to 0$ a.s. This effect is not necessarily visible for worldly sample sizes, however; log log of a (US) billion is 4.50. And as pointed out in Remark (4) below only pointwise closeness is needed.

The cited papers employ the machinery of Edgeworth-Cramér expansions and establish the validity of such using Taylor expansions of characteristic functions. Such techniques work well for fully observed i.i.d. variables but do not lend themselves easily to the present situation, due to the presence of censoring and covariates and the implicit definition of $\hat{\beta}$. There are nevertheless reasons to believe that H_n and H_n^* are sufficiently close. In the notation of earlier sections

$$\sqrt{n}\,(\hat{\beta} - \beta_o) = \{-\frac{1}{n} I(\tilde{\beta})\}^{-1} n^{-\frac{1}{2}} U(\beta_o),$$
$$\sqrt{n}\,(\hat{\beta}^* - \hat{\beta}) = \{-\frac{1}{n} I(\tilde{\beta}^*)\}^{-1} \{n^{-\frac{1}{2}} U_1^*(\hat{\beta}) + o_p(n^{-\frac{1}{2}})\},$$

where the martingales $n^{-\frac{1}{2}} U(\beta_o)$ and $n^{-\frac{1}{2}} U_1^*(\hat{\beta})$ have very similar features. They are both asymptotically normal with the same variance Σ according to Sections

2 and 3. Furthermore, with a considerable amount of effort one may show that they have skewnesses $\gamma_{1,n}/\sqrt{n}$, $\gamma_{1,n}^*/\sqrt{n}$ and kurtosises $\gamma_{2,n}/n$, $\gamma_{2,n}^*/n$ satisfying $\gamma_{1,n} \to \gamma_1$, $\gamma_{1,n}^* \to \gamma_1$, $\gamma_{2,n} \to \gamma_2$, $\gamma_{2,n}^* \to \gamma_2$ for appropriate limits γ_1, γ_2. These facts, combined perhaps with techniques as in Bhattacharya and Ghosh (1978), Hall (1983a), Withers (1983), open up for Cornish-Fisher expansion type study of the closeness of H_n^* to H_n. This is not pursued here.

(4) Let us point out the extent to which the basic assumptions (4.11) and (4.15) were used in the construction of the super bootstrap interval (4.18) and the verification of its second order correctness. (4.11) was used to get to the pivotal statement (4.13) and the α and $1 - \alpha$ points of W appearing there. It is only necessary that (4.11) holds for $t = z^{(\alpha)} - b_o/\sqrt{n} + O(n^{-1})$ and for $t = z^{(1-\alpha)} - b_o/\sqrt{n} + O(n^{-1})$, i.e. roughly only for the two points $-z^{(1-\alpha)}$ and $z^{(1-\alpha)}$. However, the O-term appearing in (4.11), or in (4.12), must be uniform for β in a neighbourhood of the reference point β_o.

(4.15) was used to get estimates of $b_o/\sqrt{n}$ and for $\hat{G}(\beta_{LOW})$, $\hat{G}(\beta_{UP})$. Precise estimates of $H_n^*(t)$ were needed only for $t = 0$ (around which the approximations work best), for $t = \{z^{(\alpha)} + b\}/\{1 - a(z^{(\alpha)} + b)\}$, and for $t = \{z^{(1-\alpha)} + b\}/\{1 - a(z^{(1-\alpha)} + b)\}$, i.e. roughly only for the three points 0, $-z^{(1-\alpha)}$, $z^{(1-\alpha)}$. But again, the almost surely statement in (4.15), which here can be read as "conditioned on $\hat{\beta}$ being very close to β_o", must be uniform over a neighbourhood of β values.

(5) The proposed confidence interval (4.18) has been given in terms of $\hat{G}^{-1}$ where $\hat{G}$ is the bootstrap distribution (1.13). Of course $\hat{G}$ is in reality only approximated as in (1.14), requiring BOOT evaluations of $\hat{\beta}^*$. The investigation of Efron (1985b, Section 8) indicates that BOOT = 1000 is a rough minimum.

(6) An important final comment in this subsection is that there are other possible approaches to second order correct intervals. If we think of the "right" interval as having endpoints $\hat{\beta} - z^{(1-\alpha)}\hat{\tau}/\sqrt{n} + A_n^{(\alpha)}/n$, $\hat{\beta} + z^{(1-\alpha)}\hat{\tau}/\sqrt{n} + A_n^{(1-\alpha)}/n$, then the problem is to get hold of the second order coefficients $A_n^{(\alpha)}$ and $A_n^{(1-\alpha)}$, and it is clear that many different methods could manage this, in the same way as there always are a variety of consistent estimators for a given statistical parameter.

The route chosen in this paper has been the one invented in Efron (1985b), but carried out in a semi-parametric model. There are also variations on the bias and acceleration corrected bootstrap interval (4.18). It may be possible to devise approximations to b_o and a_o that require less computing than the ones proposed here. Another variation could use ideas from DiCiccio and Tibshirani (1985), involving the explicit construction of a smooth transformation g having property (4.11). One such g consists of a variance stabilising mapping followed by a skewness-reducing transformation. Possessing g one could evaluate the interval (4.14) directly.

The perhaps most natural tools from classical statistics with which to fine-tune large sample results are expansions of the Edgeworth-Cramér and Cornish-Fisher type. The key idea would be to "remove skewness" in one fashion or another. (It turns out in 4.3 below that the acceleration a_o is connected to the skewness of a certain log-likelihood.) Methods using related ideas (expansion of log-likelihoods) were put forward in early important papers by Bartlett (1953a,b) and in an unpublished report by Tukey (1949). A recent reference is Abramovitch and Singh (1985) who use Edgeworth-Cramér methods to construct better intervals, both "classical" (removing skewness) and based on the bootstrap. Other references, mostly concerned with the i.i.d. nonparametric

case, are Hall (1983a) and Withers (1983).

Cox (1980), Sprott (1973, 1980), Barndorff-Nielsen (1985), and DiCiccio (1984) use saddlepoint approximation techniques to obtain second order corrected approximations to the distribution of the maximum likelihood estimator, and construct intervals based on these.

If one could effectively estimate the skewness of the distribution of $\hat{\beta}$ directly, the chi squared approximations as in Hall (1983b) would be an attractive alternative.

Still another and seemingly unrelated method uses a Bayesian framework. If one uses as endpoints the lower and upper α-point calculated from an a posteriori distribution for β, then this interval can be second order correct (in the frequentist sense adhered to here) for a cleverly chosen a priori distribution for the parameters β, $\alpha(.)$. Such an approach is investigated in general terms in Welch and Peers (1963), Welch (1965), Peers (1965), and recently extended and clarified by Stein (1985). Stein's paper is written "non-rigorously but with some care". Using such an approach in the present situation would involve constructing the a priori distribution as a solution to a differential equation (derived for the approximating finite-dimensional model also studied in 4.3 below) and then carefully checking that each in a long row of approximations involved in Stein's arguments is of sufficient precision. The a priori distribution would be improper, but different from the one derived from Jeffreys' non-informativity principle.

Let us mention a final possibility. One may invert the likelihood ratio tests for $\beta = \beta_o$, using the limiting χ^2_1 approximation. More specifically, consider

$$D_n(\beta_o) = -2 \log L(\beta_o)/L(\hat{\beta})$$
$$= 2 \sum_{i=1}^{n} \int_0^T \{(\hat{\beta} - \beta_o)z_i - \log \frac{S^{(0)}(s,\hat{\beta})}{S^{(0)}(s,\beta_o)}\} dN_i(s),$$

where L is the partial likelihood given in (1.4). One can show that $D_n(\beta_o)$ tends to a χ^2_1 in distribution under β_o, even if the likelihood only is the partial one, using results of Andersen and Gill (1982). This can of course be used to test $\beta = \beta_o$, and a natural interval is

$$I = \{\beta_o: D_n(\beta_o) \leq (z^{(1-\alpha)})^2\}.$$

That I really is an interval follows from log concavity of L.

I is a natural interval since it is based on a natural and perhaps even optimal family of tests. One can also give arguments in the direction of showing that it is second order correct in the sense discussed after (4.10). Peter Bickel has pointed out that a Bartlett correction factor could lead to even more precise intervals, with coverage probability $1 - 2\alpha + O(n^{-3/2})$.

Finding I in practice would be a difficult but solvable numerical problem. The situation is less clear when the problem is finding a second order correct confidence interval for one out of p relative risk parameters in the multi-parameter Cox model.

There is perhaps a common theme underlying the various approaches, since nearly all of them involve expansions of likelihoods in one form or another, but this theme is at present not fully understood.

4.3. Computing the acceleration constant.

The only quantity left to specify in the proposed confidence interval (4.18) is the estimate $\hat{a} = \hat{a}_o/\sqrt{n}$ of the acceleration constant $a = a_o/\sqrt{n}$. Efron (1985b) provides formulae for a in multiparameter and nonparametric models. The Cox model under study has however an infinite-dimensional nuisance parameter. Although a direct approach is possible, working only in the "continuous" Cox model, we shall below approximate it with one with only finitely many parameters, find an appropriate value for a in this model, and afterwards take a "fine limit" to get back to Cox.

The traditional statistical analysis of Cox' model starts out with the partial likelihood (1.4). The real or full likelihood can also be written down, but involves the hazard rate $\alpha(.)$:

$$L\{\beta, \alpha(.)\} = \prod_{i=1}^{n} \exp\left[\int_0^T \{\log \alpha_i(s)dN_i(s) - Y_i(s)\alpha_i(s)ds\}\right],$$

i.e.

$$\log L\{\beta, \alpha(.)\} = \sum_{i=1}^{n} \int_0^T [\{\log \alpha(s) + \beta z_i\}dN_i(s) - Y_i(s)\exp(\beta z_i)\alpha(s)ds]. \quad (4.21)$$

$L\{\beta, \alpha(.)\}$ is in fact (proportional to) the Radon-Nikodym derivative of the probability mechanism governing the (N_i, Y_i) processes under $(\beta, \alpha(.))$ w.r.t. a product of simple Poisson processes, see e.g. Brémaud and Jacod (1977, Sections 2.7 and 3.5) or Aalen (1978, Section 3).

Let us approximate the above model with one where $\alpha(.)$ is taken constant on each of many small intervals. Split [0, T] into m such intervals with mid-points s_j and lengths ds_j, and let $\alpha(s) = \alpha(s_j)$ for s in interval no. j. The resulting model has m + 1 parameters and can be handled in Efron's (1985b) framework. The intention is to let m tend to infinity and the max mesh to zero after the necessary intermediate analysis (and n is fixed). We shall, more for notational and technical convenience than out of necessity, assume that Y_i is constant on each of the m small intervals. $dN_i(s_j)$ denotes the increase of counting process N_i over interval no. j (and is 0 or 1).

First we need maximum likelihood equations and information matrix in the reduced model. Write $L^{(m)} = L^{(m)}(\theta|D)$ for the (full) likelihood in the approximating model, i.e.

$$\log L^{(m)} = \sum_{i=1}^{n} \sum_{j=1}^{m} [\{\log \alpha(s_j) + \beta z_i\} dN_i(s_j) - Y_i(s_j)\exp(\beta z_i)\alpha(s_j)ds_j], \quad (4.22)$$

where $\theta = (\beta, \alpha(s_1),\ldots,\alpha(s_m))$ is the parameter and $D = \{(N_i(s), Y_i(s)) ; s \leq T\}$ is the data. Calculations give

$$\frac{\partial \log L^{(m)}}{\partial \beta} = \sum_{i=1}^{n} \sum_{j=1}^{m} \{z_i dN_i(s_j) - Y_i(s_j) z_i \exp(\beta z_i)\alpha(s_j)ds_j\}$$

$$\doteq \sum_{i=1}^{n} \int_0^T z_i \{dN_i(s) - Y_i(s)\exp(\beta z_i)\alpha(s)ds\},$$

$$\frac{\partial \log L^{(m)}}{\partial \alpha(s_j)} = \sum_{i=1}^{n} \{dN_i(s_j)/\alpha(s_j) - Y_i(s_j)\exp(\beta z_i)\alpha(s)ds\},$$

$$= d\bar{N}(s_j)/\alpha(s_j) - S^{(0)}(s_j, \beta)ds_j;$$

$$\frac{\partial^2 \log L^{(m)}}{\partial \beta^2} = - \sum_{i=1}^{n} \sum_{j=1}^{m} Y_i(s_j) z_i^2 \exp(\beta z_i)\alpha(s_j)ds_j$$

$$\doteq - n \int_0^T S^{(2)}(s, \beta)\alpha(s)ds,$$

$$\frac{\partial^2 \log L^{(m)}}{\partial \beta \partial \alpha(s_j)} = - \sum_{i=1}^{n} Y_i(s_j) z_i \exp(\beta z_i)ds_j = - nS^{(1)}(s_j, \beta)ds_j,$$

$$\frac{\partial^2 \log L^{(m)}}{\partial \alpha(s_j)\partial \alpha(s_\ell)} = - d\bar{N}(s_j)/\alpha(s_j)^2 \, \delta_{j\ell}.$$

$\doteq$ is used to indicate the result evaluated back in the Cox model, i.e. after having m sent to infinity. It follows that the maximum likelihood estimators $\tilde{\beta}$, $\hat{\alpha}(s_1),\ldots,\hat{\alpha}(s_m)$ satisfy

$$\hat{\alpha}(s_j)ds_j = \frac{d\bar{N}(s_j)/n}{S^{(0)}(s_j, \tilde{\beta})},$$

$$\sum_{i=1}^{n} \sum_{j=1}^{m} z_i \{dN_i(s_j) - Y_i(s_j)\exp(\tilde{\beta} z_j)\hat{\alpha}(s_j)ds_j\} = 0.$$

The l.h.s. of the last equation $\doteq \sum_{i=1}^{n} \int_0^T \{z_i - E(s, \tilde{\beta})\}dN_i(s)$, where we now write

$$E(s, \beta) = S^{(1)}(s, \beta)/S^{(0)}(s, \beta). \tag{4.23}$$

It is encouraging to notice that $\tilde{\beta}$ obtained in this way, after letting m tend to infinity, is identical to the usual Cox estimator $\hat{\beta}$, cf. (1.5); indeed we shall not distinguish between them in what follows. A similar remark applies to the cumulative estimated intensity $\hat{A}$, compare (1.8). This can be taken as additional credit to the partial likelihood approach and is related to earlier findings of Johansen (1983) and Bailey (1984).

The observed information matrix is

$$\hat{I} = \begin{pmatrix} \hat{I}_{11} & \hat{I}_{12} \\ \hat{I}_{21} & \hat{I}_{22} \end{pmatrix}$$

with elements

$$\left.\begin{aligned} \hat{I}_{11} &\doteq n \int_0^T S^{(2)}(s, \hat{\beta}) d\hat{A}(s), \\ \hat{I}_{12} &= \{nS^{(1)}(s_j, \hat{\beta}) ds_j\}_{j \le m}, \\ \hat{I}_{22} &= \operatorname{diag} \{d\bar{N}(s_j)/\hat{\alpha}(s_j)^2;\ j \le m\}. \end{aligned}\right\} \tag{4.24}$$

Next we need Stein's (1956) notion of a least favourable one-parameter family for β at the fixed parameter point $\hat{\theta} = (\hat{\beta}, \hat{\alpha}(s_1), \ldots, \hat{\alpha}(s_m))$. The least favourable direction at this point is $\hat{\mu} = \hat{I}^{-1}\hat{\nu}$, where $\hat{\nu}$ is $(\partial\gamma/\partial\beta, \partial\gamma/\partial\alpha(s_1), \ldots, \partial\gamma/\partial\alpha(s_m))'$ evaluated at $\hat{\theta}$, and where $\gamma = \gamma(\theta)$ is the parameter of particular interest. Here γ is just β and

$$\hat{\mu} = \hat{I}^{-1} \begin{pmatrix} 1 \\ \\ 0 \end{pmatrix} = \begin{pmatrix} \hat{I}^{11} \\ \\ \hat{I}^{21} \end{pmatrix}. \tag{4.25}$$

Well known matrix formulae and some algebra yield

$$\hat{I}^{11} = (\hat{I}_{11} - \hat{I}_{12}\, \hat{I}_{22}^{-1}\, \hat{I}_{21})^{-1} \doteq \frac{1}{n}\, \hat{\Sigma}^{-1}, \tag{4.26}$$

$$\hat{I}^{21} = -\, \hat{I}_{22}^{-1}\, \hat{I}_{21}\, \hat{I}^{11} \doteq \{-\, E(s_j, \hat{\beta})\, \hat{\alpha}(s_j)\, \frac{1}{n}\, \hat{\Sigma}^{-1}\}_{j \le m}. \tag{4.27}$$

Stein's least favourable family is the one passing through $\hat{\theta}$ in the direction $\hat{\mu}$,

$$\hat{F} = \{L^{(m)}(\tau|\hat{D}) \equiv L^{(m)}(\hat{\theta} + \tau\hat{\mu}|\hat{D})\}. \tag{4.28}$$

τ is in a neighbourhood of 0 and is the only unknown parameter here; $\hat{D}$ denotes a hypothetical new collection of data $\{(\hat{N}_i(s), \hat{Y}_i(s)); s \leq T\}$ drawn from $L^{(m)}(\hat{\theta} + \tau\hat{\mu}|\hat{D})$. The Fisher information bound for an estimate of $\beta(\tau) = \hat{\beta} + \tau\hat{I}^{11}$, evaluated at $\tau = 0$, is the same as the bound for estimating β in the original m+1 parameter family, evaluated at $\hat{\theta}$. $\hat{F}$ is the only reduction to a one-parameter family where the problem of estimating β does not become artificially simpler. See also Tibshirani and Wasserman (1985).

Efron's formula for an estimate of the acceleration $a = a^{(m)}$ is

$$\hat{a}^{(m)} = \frac{1}{6} \text{SKEW}_{\tau=0} \{\frac{\partial}{\partial\tau} \log L^{(m)}(\hat{\theta} + \tau\hat{\mu}|\hat{D})\}. \tag{4.29}$$

The result after differentiating $\log L^{(m)}\{\hat{\beta} + \tau\hat{I}^{11}, \hat{\alpha}(s_1) - \tau\hat{E}(s_1, \hat{\beta})\hat{\alpha}(s_1)\hat{I}^{11}, \ldots, \hat{\alpha}(s_m) - \tau\hat{E}(s_m, \hat{\beta})\hat{\alpha}(s_m)\hat{I}^{11}\}$ w.r.t. τ and putting $\tau = 0$, where $\hat{E}(s_j, \hat{\beta})$ is (4.23) evaluated for $\beta = \hat{\beta}$ and with the "new data" $\hat{D}$, is the variable

$$\hat{V}^{(m)} = \hat{I}^{11} \sum_{i=1}^{n} \sum_{j=1}^{m} \{z_i - \hat{E}(s_j, \hat{\beta})\} \{d\hat{N}_i(s_j) - \hat{Y}_i(s_j)\exp(\hat{\beta}z_i)\hat{\alpha}(s_j)ds_j\}. \tag{4.30}$$

$\hat{a}^{(m)} = \frac{1}{6}$ SKEW $\{\hat{V}^{(m)}\}$ can now be evaluated in an explicit way, using ("time-discrete") ideas from the proof of the ("time-continuous") lemma in the Appendix. We are more interested in $\hat{a}$, the limit of $\hat{a}^{(m)}$ as m tends to infinity. One may prove that $\hat{a} = \frac{1}{6}$ SKEW $\{\hat{V}\}$ where $\hat{V}$ is the limit variable

$$\hat{V} = \frac{1}{n}\hat{\Sigma}^{-1} \sum_{i=1}^{n} \int_0^T \{z_i - E^*(s,\hat{\beta})\} \{dN_i^*(s) - Y_i^*(s)d\hat{A}_i(s)\}, \tag{4.31}$$

where N_i^*, Y_i^* now denote hypothetical data drawn from the distribution with cumulative hazard $\hat{A}_i$ of (1.11), i.e. exactly equivalent to coming from a bootstrap sample $(X_1^*, \delta_1^*), \ldots, (X_n^*, \delta_n^*)$, compare (1.12), (3.1). $E^*(s,\beta)$ here denotes $\Sigma_{j=1}^n z_j Y_j^*(s)\exp(\beta z_j)/\Sigma_{j=1}^n Y_j^*(s)\exp(\beta z_j) = S^{(1)*}(s,\beta)/S^{(0)*}(s,\beta)$.

One possible method of computing $\hat{a}$ is therefore to exploit the generated bootstrap data samples (even more). Each of the BOOT bootstrap samples leads not only to a realisation of $\hat{\beta}^*$ but also to a realisation of $\hat{V}$, say $\hat{V}^{*b}$. This $\hat{V}$ could for example be computed at the beginning of the numerical routine that finds $\hat{\beta}^*$, compare (3.3). Now use the empirical skewness of these BOOT $\hat{V}$-values as an estimate of the skewness of $\hat{V}$, i.e.

$$\hat{a} = \frac{1}{6} \text{ SKEW } \{V^{*1},\ldots,V^{*BOOT}\}. \tag{4.32}$$

Of course the constant $\frac{1}{n}\hat{\Sigma}^{-1}$ may be removed before computing the skewness.

The scheme above, taking advantage of the data and the model using raw computing power, is in the true bootstrap/meat axe spirit, and does not require any theoretical knowledge, however interesting, of for example the skewness of a martingale. If Statistician A followed the procedure above to fill in his value for $\hat{a} = \hat{a}_o/\sqrt{n}$ in the confidence interval (4.18), then he has done his job and has the (second order) right not to be interested in an explicit formula Statistician B may have worked out for $\hat{a}$.

The present author will nevertheless join forces with Statistician B and proceed working on alternative approaches to the computation of $\hat{a}$. Explicit (or less implicit) expressions are always valuable, and can here lead to a better understanding of the acceleration factor, but the prime reason for carrying out the reasoning below is that the formulae that are obtained are of use also in other statistical problems.

Instead of $\hat{V}$, consider

$$V = \frac{1}{n}\Sigma^{-1} \sum_{i=1}^{n} \int_0^T \{z_i - E(s,\beta_o)\} \{dN_i(s) - Y_i(s)\exp(\beta_o z_i)dA(s)\}. \tag{4.33}$$

V is just a constant times

$$U = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^T \{z_i - E(s,\beta_o)\}\, dM_i(s), \tag{4.34}$$

employing once more the martingales (2.1). $a' = \frac{1}{6}$ SKEW $\{V\}$ is considerably easier to give a formula for than is $\hat{a} = \frac{1}{6}$ SKEW $\{\hat{V}\}$, and it will also be easier to construct estimates $\hat{a}'$ for a' with sufficient precision. The results of the Appendix can be shown to imply that for the proposed $\hat{a}'$ displayed later $\hat{a}' - \hat{a} = O(n^{-1})$ a.s., or $\hat{a}_o' - \hat{a}_o = \sqrt{n}\,\hat{a}' - \sqrt{n}\,\hat{a} = O(n^{-\frac{1}{2}})$ a.s., which is good enough, compare (4.19) and the arguments following it. Using $\hat{a}'$ instead of the $\hat{a}$ that would result from explicit moment evaluations saves us from a large amount of terms like $\{1 - \Delta\hat{A}_i(s)\}^3\Delta\hat{A}_i(s) - \Delta\hat{A}_i(s)^3\{1 - \Delta\hat{A}_i(s)\}$ etc., cf. (3.8). The $\hat{a}$-ingredient chosen for citation in the preceding sentence will in its $\hat{a}'$ version be simply $\exp(\hat{\beta}z_i)d\hat{A}(s)$.

An exact expression for a' is obtained as follows, using formulae for the second and third moments to be found in the Appendix. One has $EU^2 = ER_2$ and $EU^3 = ER_3/\sqrt{n}$ where

$$R_2 = \frac{1}{n} \sum_{i=1}^{n} \int_0^T \{z_i - E(s,\beta_o)\}^2\, Y_i(s)\exp(\beta_o z_i)dA(s)$$

$$= \int_0^T \{S^{(2)}(s,\beta_o) - S^{(1)}(s,\beta_o)^2/S^{(0)}(s,\beta_o)\}dA(s), \tag{4.35}$$

$$R_3 = \frac{1}{n} \sum_{i=1}^{n} \int_0^T \{z_i - E(s,\beta_o)\}^3\, Y_i(s)\exp(\beta_o z_i)dA(s)$$

$$+ 3\,\frac{1}{n} \int_0^T \int_0^t \sum_{i=1}^{n} \{z_i - E(s,\beta_o)\}dM_i(s) \sum_{j=1}^{n} \{z_j - E(t,\beta_o)\}^2 Y_j(t)\exp(\beta_o z_j)dA(t). \tag{4.36}$$

Hence

$$a' = ER_3/\{6\sqrt{n}\,(ER_2)^{3/2}\}. \tag{4.37}$$

There are a number of ways to proceed to get hold of (a version of) $\hat{a}'$. One might stick to exact expressions, using for example

$$EY_i(s) = \Pr\{X_i^o \geq s,\ c_i \geq s\} = \exp\{-A(s)\exp(\beta_o z_i)\}\ I\{c_i \geq s\},$$

and arrive at long and complex expressions for ER_2 and ER_3. Then one should plug in $\hat{\beta}$ and $\hat{A}$ for β_o and A where necessary and compute the whole thing. (The formula above applies when the censoring time c_i is known; in the random censorship model

$$EY_i(s) = \exp\{-A(s)\exp(\beta_o z_i)\}\ R[s,\infty)$$

is appropriate, where R denotes the distribution of censoring times, and for which a Kaplan-Meier estimate is available.)

Or one might use some approximations. It is demonstrated in the Appendix that $\hat{R}_2 = \hat{\sigma}^2 = ER_2 + O_p(n^{-\frac{1}{2}})$, so that $\hat{\sigma}^2$ is a good enough estimate of EU^2, and that $\hat{R}_3 = \hat{R}_3' + \hat{R}_3''$ is within $O_p(n^{-\frac{1}{2}})$ of ER_3, where

$$\hat{R}_3' = \frac{1}{n}\sum_{i=1}^{n}\int_0^T \{z_i - E(s,\hat{\beta})\}^3\, Y_i(s)\exp(\hat{\beta}z_i)d\hat{A}(s), \tag{4.38}$$

$$\hat{R}_3'' = -3\,\frac{1}{n}\sum_{i=1}^{n}\int_0^T\int_0^t \{z_i - E(s,\hat{\beta})\}\exp(\hat{\beta}z_i)d\hat{A}(s)\{z_i - E(t,\hat{\beta})\}^2 Y_i(t)\exp(\hat{\beta}z_i)d\hat{A}(t). \tag{4.39}$$

It follows that

$$\hat{a}' = \frac{1}{6\sqrt{n}}\,\frac{\hat{R}_3' + \hat{R}_3''}{\hat{\sigma}^3} \tag{4.40}$$

is within $O_p(n^{-1})$ of $a' = \frac{1}{6}\,\mathrm{SKEW}\{U\} = \frac{1}{6}\,\mathrm{SKEW}\{V\}$, and also within $O_p(n^{-1})$ of 1/6 times the skewness of U evaluated at the maximum likelihood estimates $\hat{\beta}$ and $\hat{A}$, and finally within $O_p(n^{-1})$ of $\hat{a}$ computed as in (4.32).

The confidence interval (4.18) obtained by inserting $\hat{a}'$ above is therefore also second order correct. Variations exist. Since $\hat{\sigma}$ is computed in any case and since $\hat{R}_3'$ also is easy and natural to compute from the data, in that

it provides a measure of how skew the distribution of covariate calues z_i is in the actual experiment, one might consider keeping these in the formula for $\hat{a}'$ but bootstrapping to get hold of an equivalent version of $\hat{R}_3''$, viz.

$$\hat{R}_3''^{*} = 3\frac{1}{n}\int_0^T \int_0^t \sum_{i=1}^n \{z_i - E^*(s,\hat{\beta})\}dM_i^*(s) \sum_{j=1}^n \{z_j - E^*(t,\hat{\beta})\}^2 Y_j^*(t)\exp(\hat{\beta}z_j)d\hat{A}(t),$$

compare (4.36), (4.39).

There is basically a choice between the bootstrap based $\hat{a}$ of (4.32), requiring a large number of computations of the variable $\hat{V}$ of (4.31), and the explicit formula $\hat{a}'$ involving $\hat{R}_3'$ and $\hat{R}_3''$. The latter choice displays the oldfashioned un-bootstrap feature of consistently estimating a population parameter in an explicit way, but the first choice may still be the most practical one, given that the statistician has decided to generate bootstrap samples. A final reason for working out the $\hat{a}'$ formula is that one conceivably might construct approximations to the bootstrap distribution $\hat{G}$ itself without bootstrap samples, perhaps using Edgeworth-Cramér expansions techniques or perhaps using a method similar to one invented in DiCiccio and Tibshirani (1985).

5. Concluding remarks.

(1) It was shown in Section 4.2 how the assumptions (4.11), (4.15) led to the second order correctness of the confidence interval (4.18). General difficulties with multiparameter families caused however the somewhat evasive treatment of the acceleration factor a presented in Section 4.3, where recourse was taken to a certain least favourable one-parameter family. Thus, following Efron's (1985b) construction, we rather employed the equivalent local version of (4.11) for this least favourable reduction. The arguments of Section 4.2 easily carry over to the reduced model.

Part of the problem is the difficulty of obtaining a proper and universally agreed upon definition of a second order correct interval in the presence of nuisance parameters. We may speculate with Efron (1985b) that given a suitable and sensible definition, (4.18) will indeed be second order correct. More work is needed in this area. A study of transformations to approximate normality in multi-parameter models, parallelling and extending the work of Efron (1982b), would be welcomed.

(2) The first half of Section 4.3 was concerned with the problem of obtaining the least favourable one-parameter family in the Cox model, considering $\alpha(.)$ as an infinite-dimensional nuisance parameter. The problem was solved via intermediate analysis in a finite-dimensional approximating model. One can also obtain the same result separately and more directly, working only in the time-continuous Cox model, using a suitable extension of Stein's notion of a least favourable reduction for models with infinite-dimensional nuisance parameters. Such an analysis is indeed provided by Begun et al. (1983, Section 6) (with slightly different assumptions). Both approaches are valid; there is only the recurring problem of "where to put the hard part".

(3) The most important extension of the model studied in Sections 1-4 is the traditional p-variate Cox model, in which individual no. i has covariates $z_i = (z_{i,1},\ldots,z_{i,p})'$ and hazard rate $\alpha_i(s) = \alpha(s)\exp(\beta' z_i) = \alpha(s)\exp(\beta_1 z_{i,1} + \ldots + \beta_p z_{i,p})$. Most of the discussion of earlier sections goes through for the p-variate model, with only minor modifications. In particular the basic bootstrap scheme is as in (1.10) - (1.12), only with $\hat\beta_1 z_{i,1} + \ldots + \hat\beta_p z_{i,p}$ replacing $\hat\beta z_i$. There are now p bootstrap distributions $\hat G_i(t) = \Pr_*\{\hat\beta_i \le t\}$.

There are p-variate versions of Sections 2 and 3, involving at crucial points a multivariate martingale central limit theorem. Involved in this is a p x p covariance matrix Σ with elements

$$\sigma_{j\ell} = \int_0^T \{s_{j\ell}^{(2)}(s,\beta_o) - s_j^{(1)}(s,\beta_o)s_\ell^{(1)}(s,\beta_o)/s^{(0)}(s,\beta_o)\}dA(s), \qquad (5.1)$$

cf. (2.6), where $s_j^{(1)}(s,\beta_o)$ and $s_{j\ell}^{(2)}(s,\beta_o)$ are the limits in probability of respectively $S_j^{(1)}(s,\hat\beta) = \frac{1}{n}\Sigma_{i=1}^n z_{i,j}Y_i(s)\exp(\hat\beta' z_i)$ and $S_{j\ell}^{(2)}(s,\hat\beta) = \frac{1}{n}\Sigma_{i=1}^n z_{i,j}z_{i,\ell}Y_i(s)\exp(\hat\beta' z_i)$. Also involved is the natural analogue of (2.9), a consistent estimator $\hat\Sigma$ with elements $\hat\sigma_{j\ell}$. After working through the details one arrives at an asymptotic (first order) justification of the bootstrap procedure, and in particular it may be verified that $\hat G_i^{-1}(\alpha) \le \beta_i \le \hat G_i^{-1}(1-\alpha)$ is asymptotically equivalent to the traditional large sample theory based confidence interval for β_i, namely $\hat\beta_i - z^{(1-\alpha)}(\hat\sigma^{ii})^{\frac12}/\sqrt{n} \le \beta_i \le \hat\beta_i + z^{(1-\alpha)}(\hat\sigma^{ii})^{\frac12}/\sqrt{n}$, writing as usual $\sigma^{j\ell}$ and $\hat\sigma^{j\ell}$ for the elements of respectively Σ^{-1} and $\hat\Sigma^{-1}$.

Suppose a second order correct interval is sought for the parameter β_1. It can be demonstrated along the lines of Section 4.2 that

$$\hat G_1^{-1}\{\Phi(\hat z_1^{[\alpha]})\} \le \beta_1 \le \hat G_1^{-1}\{\Phi(\hat z_1^{[1-\alpha]})\} \qquad (5.2)$$

is such an interval, where

$$\hat{z}_1^{[1-\alpha]} = \hat{b}_1 + \frac{z^{(1-\alpha)} + \hat{b}_1}{1 - \hat{a}_1(z^{(1-\alpha)} + \hat{b}_1)} \tag{5.3}$$

and similarly for $\hat{z}_1^{[\alpha]}$, and where $\hat{b}_1 = \hat{b}_{1,o}/\sqrt{n} = \Phi^{-1}\{\hat{G}_1(\hat{\beta}_1)\}$. It remains only to find the appropriate acceleration factor $\hat{a}_1 = \hat{a}_{1,o}/\sqrt{n}$.

The treatment of this problem in the univariate case of Section 4.3 can be parallelled. The approximating finite-dimensional model with a likelihood analogous to (4.22) has p + m parameters. Skipping many details, one arrives at the equivalent of (4.31),

$$\hat{V}_1 = \frac{1}{n} \sum_{i=1}^{n} \int_0^T \hat{K}_{1,i}(s,\hat{\beta}) \{dN_i^*(s) - Y_i^*(s)d\hat{A}_i(s)\}, \tag{5.4}$$

where

$$\hat{K}_{1,i}(s,\hat{\beta}) = \sum_{u=1}^{p} \hat{\sigma}^{1u} \{z_{i,u} - E_u^*(s,\hat{\beta})\}, \tag{5.5}$$

$$E_u^*(s,\hat{\beta}) = S_u^{(1)*}(s,\hat{\beta})/S^{(0)*}(s,\hat{\beta}) = \sum_{i=1}^{n} z_{i,u} Y_i^*(s)\exp(\hat{\beta}'z_i) / \sum_{i=1}^{n} Y_i^*(s)\exp(\hat{\beta}'z_i). \tag{5.6}$$

As in Section 4.3 there are at least two ways to proceed to get hold of $\hat{a}_1 = \frac{1}{6}$ SKEW $\{\hat{V}_1\}$, or another second order equivalent version lying within $O_p(n^{-1})$ of $\hat{a}_1$. One effective method, given that the statistician has decided to generate bootstrap samples in the first place, compare the discussion in Section 4.3, is to evaluate bootstrap replicates $\hat{V}_1^{*b}$ along with the $\hat{\beta}_1^{*b}$'s, and use 1/6 times the empirical skewness of these values for $\hat{a}_1$. Another possibility is to use an explicit formula. Consider

$$V_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^T K_{1,i}(s,\beta_o) \{dN_i(s) - Y_i(s)\exp(\beta_o'z_i)dA(s)\}, \tag{5.7}$$

with

$$K_{1,i}(s,\beta_o) = \sum_{u=1}^{p} \sigma^{1u} \{z_{i,u} - E_u(s,\beta_o)\}, \tag{5.8}$$

writing $E_u(s,\beta) = S_u^{(1)}(s,\beta)/S^{(0)}(s,\beta)$. The plan is to evaluate the skewness of V_1 at the maximum likelihood estimators $\hat{\beta}$, $\hat{A}$.

Some matrix algebra combined with results from the Appendix can be used to show that $EV_1^2 = \sigma^{11} + O(n^{-\frac{1}{2}}) = \hat{\sigma}^{11} + O_p(n^{-\frac{1}{2}})$, i.e. $\hat{\sigma}^{11} = EV_1^2 + O_p(n^{-\frac{1}{2}})$. Also needed is EV_1^3, expressions for which are obtainable via the methods of the Appendix. The result is that

$$\hat{a}_1' = \frac{1}{6\sqrt{n}} \frac{\hat{R}_{1,3}' + \hat{R}_{1,3}''}{(\hat{\sigma}^{11})^{3/2}} = \hat{a}_1 + O_p(n^{-1}), \tag{5.9}$$

where

$$\hat{R}_{1,3}' = \frac{1}{n} \sum_{i=1}^{n} \int_0^T [\sum_{u=1}^{p} \hat{\sigma}^{1u}\{z_{i,u} - E_u(s,\hat{\beta})\}]^3 Y_i(s)\exp(\hat{\beta}'z_i)d\hat{A}(s), \tag{5.10}$$

$$\hat{R}_{1,3}'' = -3\frac{1}{n} \sum_{i=1}^{n} \int_0^T \int_0^t [\sum_{u=1}^{p} \hat{\sigma}^{1u}\{z_{i,u} - E_u(s,\hat{\beta})\}]\exp(\hat{\beta}'z_i)d\hat{A}(s)$$

$$[\sum_{u=1}^{p} \hat{\sigma}^{1u}\{z_{i,u} - E_u(t,\hat{\beta})\}]^2\exp(\hat{\beta}'z_i)Y_i(t)d\hat{A}(t). \tag{5.11}$$

(4) The formulae derived in the Appendix for the skewness of a martingale have use also in other problems. Let us briefly point out two applications.

(4a) Consider first a <u>parametric</u> Cox model, where the underlying common $\alpha(.)$ is assumed constant over the time interval $[0, T]$, say $\alpha_i(s) = \theta\exp(\beta z_i)$. The log likelihood for this model, provided the data are collected over this interval, for example in the form of counting processes N_i and at-risk indicators Y_i as in previous sections, becomes

$$\log L(\beta,\theta) = \sum_{i=1}^{n} \int_0^T \{(\log\theta + \beta z_i)dN_i(s) - Y_i(s)\theta\exp(\beta z_i)ds\}. \tag{5.12}$$

The maximum likelihood estimators $\hat{\beta}$, $\hat{\theta}$ solve the equations

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} \int_0^T z_i\{dN_i(s) - Y_i(s)\theta\exp(\beta z_i)ds\} = 0,$$

$$\frac{\partial \log L}{\partial \theta} = \sum_{i=1}^{n} \int_0^T \frac{1}{\theta}\{dN_i(s) - Y_i(s)\theta\exp(\beta z_i)ds\} = 0.$$

$\hat{\beta}$ and $\hat{\theta}$ are asymptotically optimal estimators by the results of Hjort (1985).

Bootstrapping can be performed as follows: Generate X_i^{o*} from the estimated distribution $\hat{F}_i$, exponential with parameter $\hat{\theta}\exp(\hat{\beta}z_i)$, and let $X_i^* = \min\{X_i^{o*}, c_i\}$, $\delta_i^* = I\{X_i^{o*} \le c_i\}$. This leads to N_i^* and Y_i^* as in (3.3), $i = 1,\ldots,n$, and then to bootstrap replicates $\hat{\beta}^*$, $\hat{\theta}^*$. Let $\hat{G}(t) = \Pr_*\{\hat{\beta}^* \le t\}$.

Suppose a second order correct confidence interval for β is to be constructed. This can be done exactly as in Section 4, but with a new formula for $\hat{a} = \hat{a}_o/\sqrt{n}$. Following the reasoning of Section 4 in this case, which is similar but in fact much easier, one gets $\hat{a} = \frac{1}{6}\,\mathrm{SKEW}\,\{\hat{V}\}$, where

$$\hat{V} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^T \{z_i - E(\hat{\beta})\}\,\{dN_i^*(s) - Y_i^*(s)\hat{\theta}\exp(\hat{\beta}z_i)ds\}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{z_i - E(\hat{\beta})\}\,\{N_i^*(T) - \hat{\theta}\exp(\hat{\beta}z_i) \int_0^T Y_i^*(s)ds\}. \qquad (5.13)$$

Here $E(\beta) = \int_0^T S^{(1)}(s,\beta)ds / \int_0^T S^{(0)}(s,\beta)ds$. $\hat{a}$ may be computed using bootstrap replications of $\hat{V}$, or using

$$E\hat{V}^2 \doteq \hat{\theta}\,\{F^{(2)}(\hat{\beta}) - F^{(1)}(\hat{\beta})^2/F^{(0)}(\hat{\beta})\},$$

$$E\hat{V}^3 \doteq \frac{1}{\sqrt{n}}\Big[\frac{1}{n}\sum_{i=1}^{n}\{z_i - E(\hat{\beta})\}^3\hat{\theta}\exp(\hat{\beta}z_i)\int_0^T Y_i(s)ds$$

$$- 3\,\frac{1}{n}\sum_{i=1}^{n}\{z_i - E(\hat{\beta})\}^3\hat{\theta}^2\exp(2\hat{\beta}z_i)\int_0^T tY_i(t)dt\Big],$$

where $F^{(k)}(\beta) = \int_0^T S^{(k)}(s,\beta)ds$.

Similarly other first order asymptotic statements in other parametric survival data models may be "corrected" to second order using the bootstrap approach and results from this paper.

(4b) As a second example of the use of the developed methods, consider the "Kaplan-Meier problem" of obtaining nonparametric estimates of the unknown cumulative distribution F and the unknown cumulative hazard $A(t) = \int_0^t \alpha(s)ds$ based on a partially censored sample of n observations from F. Let again X_i^o be the uncensored observation for individual no. i and let c_i be the censoring time, so that $X_i = \min\{X_i^o, c_i\}$ and $\delta_i = I\{X_i^o \leq c_i\}$ are observed.

The natural estimator for A is Nelson and Aalen's $\hat{A}(t) = \int_0^t d\bar{N}(s)/\bar{Y}(s)$, where $N_i(s) = I\{X_i^o \leq s, \delta_i = 1\}$, $Y_i(s) = I\{X_i^o \geq s, c_i \geq s\}$, and $\bar{N} = \Sigma_{i=1}^n N_i$, $\bar{Y} = \Sigma_{i=1}^n Y_i$. The Kaplan-Meier estimator for F is $\hat{F}(t) = 1 - \Pi_{[0,t]}\{1 - d\bar{N}(s)/\bar{Y}(s)\}$, and lies uniformly within $O_p(n^{-1})$ of $1 - \exp\{-\hat{A}(t)\}$, see for example Hjort (1984, Section 3).

If $M_i(t) = N_i(t) - \int_0^t Y_i(s)\alpha(s)ds$ and $\bar{M} = \Sigma_{i=1}^n M_i$, and $\bar{A}(t) = \int_0^t \bar{J}(s)\alpha(s)ds$ where $\bar{J}(s) = I\{\bar{Y}(s) > 0\}$, then

$$Z_n(t) = \sqrt{n}\,\{\hat{A}(t) - \bar{A}(t)\} = \sqrt{n} \int_0^t \bar{J}(s)d\bar{M}(s)/\bar{Y}(s)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t \{n\bar{J}(s)/\bar{Y}(s)\}\, dM_i(s). \qquad (5.14)$$

Let us assume the random censorship model, where the c_i's are drawn independently from a distribution R, and independently of the X_i^o's. Then $EY_i(s) = F[s,\infty)R[s,\infty) = \exp\{-A(s)\}R[s,\infty) = y(s)$, and $\bar{Y}(s)/n$ converges to y(s) uniformly on [0, T] in probability, also $n\bar{J}(s)/\bar{Y}(s)$ converges to $1/y(s)$ uniformly on [0, T] in probability provided y(s) is bounded away from zero in this interval (which amounts to saying $y(T) > 0$).

Z_n is a zero mean martingale, and

$$EZ_n(t)^2 = E\,\frac{1}{n} \sum_{i=1}^n \int_0^t \{n\bar{J}(s)/\bar{Y}(s)\}^2 Y_i(s)\alpha(s)ds$$

$$= E \int_0^t \{n\bar{J}(s)/\bar{Y}(s)\}\alpha(s)ds \;\to\; \int_0^t \{1/y(s)\}\alpha(s)ds = \sigma(t)^2. \qquad (5.15)$$

It is a well known fact by now that $Z_n(t) \xrightarrow{d} Z(t)$, a Gaussian martingale with Var $Z(t) = \sigma(t)^2$. This result combined with a natural estimator of $\sigma(t)$ is widely used to get pointwise or simultaneous confidence bands for $A(t)$ and a fortiori (via $F(t) = 1 - \exp\{-A(t)\}$) for $F(t)$, see for example Andersen and Borgan (1985) and Hjort's (1985a) discussion contribution.

The methods and results of the present paper make it possible to obtain the asymptotic skewness of the Nelson-Aalen estimator. The formulae in A.III can be shown to imply

$$\begin{aligned} \text{SKEW}\{\hat{A}(t)\} &= \text{SKEW}\{Z_n(t)\} \\ &= \frac{1}{\sqrt{n}} \Big[\int_0^t \{1/y(s)^2\}\alpha(s)ds + \frac{3}{2}\sigma(t)^4 \Big] \Big/ \sigma(t)^3 + O(n^{-1}). \end{aligned} \quad (5.16)$$

The slightly alarming consequence is that the distribution of $\hat{A}(t)$ always is skewed to the right (a lower bound for the skewness is $\{A(t)^{-1} + \frac{3}{2}\}\sigma(t)/\sqrt{n} + O(n^{-1})$, using the Cauchy-Schwarz inequality). The traditional first order asymptotic statements effectively ignore this positive skewness. There are ways of "repairing for skewness" now that it has been detected. These matters will perhaps be returned to at a later occasion.

Acknowledgements. This paper has been written during my stay at Stanford University with grants from the Norwegian Computing Center and the Royal Norwegian Council for Scientific and Industrial Research. Comments from Rudolf Beran, Peter Bickel, Bradley Efron, Tore Schweder, and Charles Stein have been helpful.

APPENDIX

A.I. Preliminaries.

Some of the more involved technical arguments that were needed several places in Sections 2-4 are tended to here. We shall not hesitate to make convenient assumptions as long as they are statistically reasonable. Thus we postulate at once that the covariates are uniformly bounded, i.e.

$$|z_i| \leq M \text{ for every } z_i \tag{A.1}$$

for some large enough M. Following Andersen and Gill (1982) we assume throughout the Appendix that $S^{(k)}(s,\beta) = \frac{1}{n}\Sigma_{j=1}^{n}(z_j)^k Y_j(s)\exp(\beta z_j)$ converges to some $s^{(k)}(s,\beta)$ uniformly for s in [0, T] and for β in a neighbourhood of each reference point β_o, in probability, k = 0, 1, 2, 3. This is indeed a consequence of the reasonable assumption that the empirical distribution of the first n z_i's converges to some distribution on [-M, M] and that the empirical distribution of the first n c_i's converges to some distribution R on [0, ∞] with the property that R[T, ∞] > 0. (No censoring at all corresponds to R{∞} = 1.) The limit functions $s^{(k)}(s,\beta)$ are continuous in β, so that $\sup_{t\leq T} |S^{(k)}(s,\hat{\beta}) - s^{(k)}(s,\beta_o)| \xrightarrow{p} 0$ too. Finally $s^{(0)}(s,\beta)$ is bounded away from zero for $s \leq T$ and for β in a neighbourhood of β_o.

Since we aim at second order asymptotic results we must ask for even more than mentioned above. Define

$$\begin{aligned} C_n(t) &= n^{\frac{1}{2}}\{S^{(k)}(t,\beta) - s^{(k)}(t,\beta)\} \\ &= n^{\frac{1}{2}}\{S^{(k)}(t,\beta) - ES^{(k)}(t,\beta)\} + n^{\frac{1}{2}}\{\frac{1}{n}\sum_{j=1}^{n}(z_j)^k y_j(t)\exp(\beta z_j) - s^{(k)}(t,\beta)\} \\ &= C_{n,1}(t) + C_{n,2}(t), \end{aligned} \tag{A.2}$$

writing $y_j(t) = EY_j(t) = \Pr\{X_j^o \geq t, c_j \geq t\}$. Assume for the moment that the c_j's are drawn independently from the distribution R; the case with known censoring times can be handled similarly but with slightly heavier notation.

A fair amount of details lead to

$$E|C_{n,1}(t) - C_{n,1}(s)|^2\ |C_{n,1}(u) - C_{n,1}(t)|^2$$

$$\le \frac{1}{n}\sum_{j=1}^{n}(z_j)^{2k}\exp(2\beta z_j)\delta_j(s,t)\ \frac{1}{n}\sum_{j=1}^{n}(z_j)^{2k}\exp(2\beta z_j)\delta_j(t,u)$$

$$+ 2\left\{\frac{1}{n}\sum_{j=1}^{n}(z_j)^{2k}\exp(2\beta z_j)\ \delta_j(s,t)\delta_j(t,u)\right\}^2,\ s \le t \le u,$$

where $\delta_j(s,t) = y_j(s) - y_j(t) = F_j[s,\infty)R[s,\infty) - F_j[t,\infty)R[t,\infty) \le F_j[s,t) + R[s,t)$ $\le \exp(\beta_o z_j)\{A(t) - A(s)\} + R[s,t)$. Hence the above expectation is

$$\le 3\ M^{2k}\exp(2|\beta|M)\ [\exp(|\beta|M)\{A(t)-A(s)\} + R[s,t)]$$

$$M^{2k}\exp(2|\beta|M)\ [\exp(|\beta|M)\{A(u)-A(t)\} + R[t,u)].$$

This can be used to prove that $\{C_{n,1}\}$ is tight, following the arguments in the proof of Theorem 15.6 in Billingsley (1968). Similarly $\{C_{n,2}\}$ is tight. All of the arguments used can be made uniform over bounded sets of values of β. The conclusion we need from all this is

$$\sup_{\beta\in B_o}\ \sup_{0\le t\le T}\ |S^{(k)}(t,\beta) - s^{(k)}(t,\beta)| = O_p(n^{-\frac{1}{2}}). \tag{A.3}$$

In fact $C_n(.)$ will converge to a zero mean Gaussian process in $D[0, T]$. A final consequence of the efforts above is

$$\sup_{0\le t\le T}\ |S^{(k)}(t,\hat{\beta}) - s^{(k)}(t,\beta_o)| = O_p(n^{-\frac{1}{2}}). \tag{A.4}$$

A.II. The convergence of $\int H_n(s)d\hat{A}(s)$.

The lemma stated as (3.9) was needed several times in the course of Section 3 and is also needed, along with an error estimate, in connection with the problem of estimating the accerelation factor $a = a_o/\sqrt{n}$ with sufficient precision. The proof below is some steps longer than actually needed to show convergence in probability since we in some cases shall need to know that the convergence rate is $n^{\frac{1}{2}}$.

Lemma A.1. Let $H_n(.)$ be a predictable and bounded stochastic process, and assume that H_n converges to some h uniformly in probability. Then $\int_0^T H_n(s)d\hat{A}(s) \xrightarrow{p} \int_0^T h(s)dA(s)$. Furthermore, if $\sup_{0\le s\le T} |H_n(s) - h(s)| = O_p(n^{-\frac{1}{2}})$, then $\sup_{0\le t\le T}|\int_0^t H_n(s)d\hat{A}(s) - \int_0^t h(s)dA(s)| = O_p(n^{-\frac{1}{2}})$.

Proof: Writing $\bar{N} = \Sigma_{i=1}^n N_i$, $\bar{M} = \Sigma_{i=1}^n M_i$, we deduce from $dN_i(s) = dM_i(s) + Y_i(s)\exp(\beta_o z_i)dA(s)$ that $d\bar{N}(s)/n = d\bar{M}(s)/n + S^{(0)}(s,\beta_o)dA(s)$. Hence by a Taylor expansion argument

$$\int_0^T H_n(s)d\hat{A}(s) = \int_0^T H_n(s)\frac{d\bar{N}(s)/n}{S^{(0)}(s,\hat{\beta})}$$

$$= \int_0^T H_n(s)\{\frac{1}{S^{(0)}(s,\beta_o)} - \frac{S^{(1)}(s,\beta_o)}{S^{(0)}(s,\beta_o)^2}(\hat{\beta}-\beta_o) + \tfrac{1}{2}\frac{\partial^2}{\partial\beta^2}\frac{1}{S^{(0)}(s,\tilde{\beta})}(\hat{\beta}-\beta_o)^2\}\frac{d\bar{M}(s)}{n}$$

$$+ \int_0^T H_n(s)\{1 - \frac{S^{(1)}(s,\beta_o)}{S^{(0)}(s,\beta_o)}(\hat{\beta}-\beta_o) + \tfrac{1}{2}\frac{\partial^2}{\partial\beta^2}\frac{1}{S^{(0)}(s,\tilde{\beta})}(\hat{\beta}-\beta_o)^2 S^{(0)}(s,\beta_o)\}dA(s)$$

$$= (i) + (ii) + (iii) + (iv) + (v) + (vi),$$

say, with $\tilde{\beta}$ somewhere between β_o and $\hat{\beta}$.

(i) is the martingale $L(t) = n^{-\frac{1}{2}}\int_0^t H_n(s)S^{(0)}(s,\beta_o)^{-1}d\bar{M}(s)$ evaluated at T and divided by $\sqrt{n}$. The martingale has variance process

$$(L,L)(t) = \frac{1}{n}\sum_{i=1}^n \int_0^t H_n(s)^2 S^{(0)}(s,\beta_o)^{-2} Y_i(s)\exp(\beta_o z_i)dA(s)$$

$$= \int_0^t H_n(s)^2 S^{(0)}(s,\beta_o)^{-1}dA(s).$$

Lenglart's inequality, see for example Andersen and Borgan (1985, Section 3), implies

$$\Pr\{\sup_{0\le t\le T}|L(t)| > \eta\} \le \delta/\eta^2 + \Pr\{(L,L)(T) > \delta\}$$

for all positive δ and η. Choosing δ big enough to make the second term less than

a prescribed ε and afterwards an even bigger n to get δ/η^2 less than ε too, we see that $\sup_{0\le t\le T} |L(t)|$, and in particular $L(T)$, is bounded in probability. Consequently (i) = $O_p(n^{-\frac{1}{2}})$.

(ii) is similarly $(\hat\beta - \beta_o)$ times $O_p(n^{-\frac{1}{2}})$, i.e. (ii) = $O_p(n^{-1})$. (iii) is $(\hat\beta - \beta_o)^2$ times $\int_0^T K_n(s)d\bar M(s)/n$, say, where K_n is bounded in probability. But $|\int_0^T K_n(s)d\bar M(s)/n| \le \frac{1}{n}\sum_{i=1}^n \int_0^T |K_n(s)|\{dN_i(s) + Y_i(s)\exp(\beta_o z_i)dA(s)\} \le \{1 + A(T)\frac{1}{n}\sum_{i=1}^n \exp(\beta_o z_i)\}\sup_{0\le s\le T}|K_n(s)|$. Hence (iii) = $O_p(n^{-1})$.

(iv) is of course within $|\int_0^T \{H_n(s) - h(s)\}dA(s)| \le A(T)\sup_{0\le s\le T}|H_n(s) - h(s)|$ of the limit $\int_0^T h(s)dA(s)$, and is accordingly $O_p(n^{-\frac{1}{2}})$ under the extra assumption stated in the lemma. (v) = $O_p(n^{-\frac{1}{2}})$ for reasons similar to those explained in connection with (iii). Finally (vi) = $O_p(n^{-1})$. □

There are also occasions where $\int_0^T H_n(s)d\hat A(s)$ needs to be studied for processes H_n that are not predictable. The most important of these cases are of the type $H_n(s) = H_n(s,\hat\beta)$, say, where $H_n(s,\beta_o)$ <u>is</u> predictable and converges uniformly in probability to some $h(s,\beta_o)$. ($H_n(s,\hat\beta)$ is <u>not</u> predictable since it depends upon $\hat\beta$ which is not even measurable w.r.t. the history up to time s.) Then another Taylor expansion saves the day:

$$\int_0^T H_n(s,\hat\beta)d\hat A(s) = \int_0^T \{H_n(s,\beta_o) + \frac{\partial}{\partial\beta}H_n(s,\beta_o)(\hat\beta - \beta_o) + \frac{1}{2}\frac{\partial^2}{\partial\beta^2}H_n(s,\tilde\beta)(\hat\beta - \beta_o)^2\}d\hat A(s),$$

and these terms can be shown to be $\int_0^T h(s,\beta_o)dA(s) + O_p(n^{-\frac{1}{2}}) + O_p(n^{-\frac{1}{2}}) + O_p(n^{-1})$ under reasonable restrictions.

A.III. The skewness of a martingale.

To arrive at suitable expressions for the acceleration factor $a = a_o/\sqrt{n}$ we needed the skewness of the random variable

$$U = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^T \{z_i - E(s,\beta_o)\} \{dN_i(s) - Y_i(s)\exp(\beta_o z_i)\alpha(s)ds\}. \qquad (A.5)$$

The formulae derived below are also useful in other statistical problems. They can be used to obtain estimates of the skewness of parameter estimators in models with censoring, pointing to the possibility of correcting first order asymptotic statements in such situations, and can also be used to obtain the skewness of nonparametric estimators like the Nelson-Aalen and the Kaplan-Meier estimators.

Let in general terms $N_1,\ldots,N_n$ be 0-1 counting processes observed over the time interval [0, T] with intensities of the form $Y_1(s)dA_1(s),\ldots,Y_n(s)dA_n(s)$, i.e.

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)dA_i(s), \; t \geq 0, \quad i = 1,\ldots,n \qquad (A.6)$$

become martingales w.r.t. the σ-algebras $F_t = \sigma\{N_i(s), Y_i(s); s \leq t\}$. We take the at-risk indicators Y_i to be non-increasing left continuous 0-1 processes, and take the martingales to be orthogonal, i.e. M_iM_j is a martingale when $i \neq j$. Finally let $dA_i(s) = \alpha_i(s)ds$, i.e. A_i is absolutely continuous, $i = 1,\ldots,n$.

Lemma A.2. Let $h_1,\ldots,h_n$ be a.s. bounded and predictable processes. (It suffices for h_i to be a caglad function, i.e. the paths are left continuous with right hand limits, and to be progressively measurable w.r.t. the F_t's. The interpretation is that $h_i(s)$ is known already at time $s-\varepsilon$ for small enough ε.) Consider

$$Z_i = \int_0^T h_i(s)dM_i(s), \quad i = 1,\ldots,n. \qquad (A.7)$$

Then $Z_1,\ldots,Z_n$ are orthogonal and the following formulae are true:

$$E\Big(\sum_{i=1}^{n} Z_i\Big)^2 = E\sum_{i=1}^{n}\int_0^T h_i(s)^2 Y_i(s)\,dA_i(s), \tag{A.8}$$

$$E(Z_i^3) = E\int_0^T h_i(s)^3 Y_i(s)\,dA_i(s) - 3\,E\int_0^T\int_0^t h_i(s)\,dA_i(s)\; h_i(t)^2 Y_i(t)\,dA_i(t), \tag{A.9}$$

$$E(Z_i Z_j^2) = E\int_0^T\int_0^t h_i(s)\,dM_i(s)\; h_j(t)^2 Y_j(t)\,dA_j(t),\ i \neq j, \tag{A.10}$$

$$E\Big(\sum_{i=1}^{n} Z_i\Big)^3 = E\sum_{i=1}^{n}\int_0^T h_i(s)^3 Y_i(s)\,dA_i(s)$$

$$+ 3\,E\int_0^T\int_0^t \sum_{i=1}^{n} h_i(s)\,dM_i(s)\;\sum_{j=1}^{n} h_j(t)^2 Y_j(t)\,dA_j(t). \tag{A.11}$$

Proof: h_i may be a function of N_j and Y_j values for $j \neq i$, as is the case for U of (A.5). Z_i and Z_j may accordingly be dependent, but they are nevertheless orthogonal martingales (here evaluated at the endpoint T). This and formula (A.8) are standard facts, see for example Andersen and Borgan (1985) or Gill (1980).

Think of Z_i as a Lebesgue-Stieltjes integral,

$$Z_i \doteq Z_i' = \sum_u h_i(s_u)\,dM_i(s_u),$$

where $dM_i(s_u) = dN_i(s_u) - Y_i(s_u)\,dA_i(s_u) = N_i[s_u, s_{u+1}) - Y_i(s_u)A_i[s_u, s_{u+1}) \doteq M_i(s_{u+1}) - M_i(s_u)$, for a fine grid $0 = s_o < \ldots < s_m = T$. Given the history of everything happened in $[0, s)$, i.e. $\mathcal{F}_{s-}$, $Y_i(s)$ is known and $N_i[s, s+ds)$ is binomial $\{Y_i(s),\ A_i[s, s+ds)\}$. This can be used to evaluate the expectation of

$$(Z_i')^3 = \sum_u h_i(s_u)^3 dM_i(s_u)^3 + 3\sum_{u<v} h_i(s_u)\,dM_i(s_u)\; h_i(s_v)^2 dM_i(s_v)^2$$

$$+ 3\sum_{u<v} h_i(s_u)^2 dM_i(s_u)^2\; h_i(s_v)\,dM_i(s_v)$$

$$+ 6\sum_{u<v<w} h_i(s_u)\,dM_i(s_u)\,h_j(s_v)\,dM_i(s_v)\,h_i(s_w)\,dM_i(s_w).$$

Here $E\{h_i(s_u)^3 dM_i(s_u)^3 | F_{u-}\} = h_i(s_u)^3 Y_i(s_u)\ [\{1-dA_i(s_u)\}^3 dA_i(s_u) - dA_i(s_u)^3 \{1-dA_i(s_u)\}] \doteq h_i(s_u)^3 Y_i(s_u) dA_i(s_u)$,

$$E\{h_i(s_u) dM_i(s_u) h_i(s_v)^2 dM_i(s_v)^2 | F_{v-}\}$$

$$= h_i(s_u) dM_i(s_u) h_i(s_v)^2 Y_i(s_v)\ [\{1-dA_i(s_v)\}^2 dA_i(s_v) + dA_i(s_v)^2 \{1-dA_i(s_v)\}]$$

$$\doteq h_i(s_u) dM_i(s_u) h_i(s_v)^2 Y_i(s_v) dA_i(s_v)$$

$$= - h_i(s_u) dA_i(s_u) h_i(s_v)^2 Y_i(s_v) dA_i(s_v),$$

and similarly $E\{h_i(s_u)^2 dM_i(s_u)^2 h_i(s_v) dM_i(s_v) | F_{v-}\} = 0$, $E\{h_i(s_u) dM_i(s_u) h_i(s_v) dM_i(s_v) h_i(s_w) dM_i(s_w) | F_{w-}\} = 0$. Hence

$$E(Z_i')^3 = E \sum_u h_i(s_u)^3 Y_i(s_u) dA_i(s_u) - 3 E \sum_{u<v} h_i(s_u) dA_i(s_u) h_i(s_v)^2 Y_i(s_v) dA_i(s_v).$$

Formula (A.9) follows by appropriate limiting arguments. Rigor can be achieved by first proving the formula for h_i a predictable step function and next for a general a.s. bounded predictable process, compare the abstract way in which such processes are defined in Meyer (1971).

Next look at

$$(\sum_{i=1}^n Z_i)^3 = \sum_{i=1}^n Z_i^3 + 3 \sum_{i \neq j} Z_i Z_j^2 + 6 \sum_{i<j<k} Z_i Z_j Z_k.$$

Using arguments similar to those above formula (A.10) can be proved, whereas $E(Z_i Z_j Z_k) = 0$ when the indices are distinct. It follows that

$$E(\sum_{i=1}^n Z_i)^3 = E \{ \sum_{i=1}^n \int_0^T h_i(s)^3 Y_i(s) dA_i(s) - 3 \sum_{i=1}^n \int_0^T \int_0^t h_i(s) dA_i(s) h_i(t)^2 Y_i(t) dA_i(t)$$

$$+ 3 \sum_{i \neq j} \int_0^T \int_0^t h_i(s) dM_i(s)\ h_j(t)^2 Y_j(t) dA_j(t) \},$$

from which the final formula (A.11) follows upon using $dM_i(s) Y_i(t) = - dA_i(s) Y_i(t)$. □

Now let us apply these formulae to the variable U of (A.5) and (4.33). Firstly, $EU^2 = ER_2$, where

$$R_2 = \frac{1}{n} \sum_{i=1}^{n} \int_0^T \{z_i - E(s,\beta_o)\}^2 Y_i(s)\exp(\beta_o z_i)dA(s)$$

$$= \int_0^T \{S^{(2)}(s,\beta_o) - S^{(1)}(s,\beta_o)^2/S^{(1)}(s,\beta_o)\}dA(s).$$

Clearly R_2 converges in probability to an underlying population parameter ρ_2 which is exactly $\Sigma = \sigma^2$, compare (2.6). The natural estimator $\hat{R}_2$ obtained by inserting $\hat{\beta}$ and $\hat{A}$ for β_o and A where necessary is just $\hat{\Sigma} = \hat{\sigma}^2$ of (2.9). The assumptions and results of A.I imply $ER_2 = \sigma^2 + O(n^{-\frac{1}{2}})$, $\hat{R}_2 = \hat{\sigma}^2 = \sigma^2 + O_p(n^{-\frac{1}{2}})$, and hence $\hat{\sigma}^2 = ER_2 + O_p(n^{-\frac{1}{2}})$, i.e. $\hat{\sigma}^2$ is a good enough estimate of EU^2.

Secondly, $EU^3 = ER_3/\sqrt{n}$, where

$$R_3 = \frac{1}{n} \sum_{i=1}^{n} \int_0^T \{z_i - E(s,\beta_o)\}^3 Y_i(s)\exp(\beta_o z_i)dA(s)$$

$$+ 3\,\frac{1}{n} \int_0^T \int_0^t \sum_{i=1}^{n} \{z_i - E(s,\beta_o)\}dM_i(s) \sum_{j=1}^{n} \{z_j - E(t,\beta_o)\}^2 Y_j(t)\exp(\beta_o z_j)dA(t)$$

$$= R_3' + R_3'' \,. \qquad \text{(A.12)}$$

The following considerations will show that there are certain population parameters ρ_3', ρ_3'' such that

$$ER_3' = \rho_3' + O(n^{-\frac{1}{2}}), \qquad ER_3'' = \rho_3'' + O(n^{-\frac{1}{2}}), \qquad \text{(A.13)}$$

so that

$$EU^3 = \frac{1}{\sqrt{n}} (\rho_3' + \rho_3'') + O(n^{-1}). \qquad \text{(A.14)}$$

ρ_3' and ρ_3'' can be expressed in terms of the limit functions $s^{(k)}(s,\beta)$, k = 0, 1, 2, 3; in fact and for the record

$$\rho_3' = \int_0^T g'(s,\beta_o)dA(s), \qquad \rho_3'' = -3 \int_0^T g''(s,\beta_o)dA(s), \qquad \text{(A.15)}$$

where

$$g_3'(s,\beta_o) = \text{p-limit of } \frac{1}{n}\sum_{i=1}^{n}\{z_i - E(s,\beta_o)\}^3\, Y_i(s)\exp(\beta_o z_i), \qquad (A.16)$$

$$g_3''(t,\beta_o) = \text{p-limit of } \frac{1}{n}\sum_{i=1}^{n}\int_0^t \{z_i - E(s,\beta_o)\}\exp(\beta_o z_i)\,dA(s)\{z_i - E(t,\beta_o)\}^2\exp(\beta_o z_i). \qquad (A.17)$$

The specific form of these population parameters need not concern us here, what is important is to get $n^{\frac{1}{2}}$-consistent estimates of them.

ρ_3', clearly the limit in probability of R_3', is easy to handle. The natural estimator is

$$\hat{R}_3' = \frac{1}{n}\sum_{i=1}^{n}\int_0^T \{z_i - E(s,\hat{\beta})\}^3\, Y_i(s)\exp(\hat{\beta} z_i)\,d\hat{A}(s). \qquad (A.18)$$

The results of A.I and A.II imply $ER_3' = \rho_3' + O(n^{-\frac{1}{2}})$, $\hat{R}_3' = \rho_3' + O_p(n^{-\frac{1}{2}})$, so $\hat{R}_3' = ER_3' + O_p(n^{-\frac{1}{2}})$ is good enough for ER_3'. Next consider ER_3'' and ρ_3''. Write

$$U_n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t \{z_i - E(s,\beta_o)\}\, dM_i(s),$$

$$B_n(t) = \sqrt{n}\,\Big[\frac{1}{n}\sum_{j=1}^{n}\{z_j - E(t,\beta_o)\}^2 Y_j(t)\exp(\beta_o z_j) - \{s^{(2)}(t,\beta_o) - s^{(1)}(t,\beta_o)^2/s^{(0)}(t,\beta_o)\}\Big],$$

so that

$$ER_3'' = 3\,E\int_0^T U_n(t)B_n(t)\,dA(t). \qquad (A.19)$$

Here $\{U_n(t), B_n(t)\}$ will in fact converge jointly to some Gaussian $\{U(t), B(t)\}$, $\int_0^T U_n(t)B_n(t)\,dA(t)$ will converge in distribution to $\int_0^T U(t)B(t)\,dA(t)$, and

$$ER_3'' \to \rho_3'' = 3\int_0^T EU(t)B(t)\,dA(t) \qquad (A.20)$$

gives another interpretation of the population parameter ρ_3''.

We shall have to be somewhat circumventive now, due to the fact that although $EdM_i(s)Y_j(t) = 0$ for $i \neq j$, $EdM_i(s)\{z_j - E(t,\beta_o)\}^2 Y_j(t)$ may still be different from zero. If $B_n^o(t)$ is as $B_n(t)$ above, but with the deterministic

limit function $e(t,\beta_o) = s^{(1)}(t,\beta_o)/s^{(0)}(t,\beta_o)$ replacing $E(t,\beta_o)$, then

$$B_n(t) - B_n^o(t) = -\sqrt{n}\,\{E(t,\beta_o) - e(t,\beta_o)\}^2\, S^{(0)}(t,\beta_o),$$

and $\sup_{t\le T} |B_n(t) - B_n^o(t)| = O_p(n^{-\frac{1}{2}})$ according to A.I. Using this one may show that

$$ER_3'' = 3\,E\int_0^T U_n(t)B_n^o(t)\,dA(t) + O(n^{-\frac{1}{2}}),$$

compare (A.19), and this latter integral is easier to handle since $E\{z_i - E(s,\beta_o)\}$ $dM_i(s)$ $\{z_j - e(t,\beta_o)\}^2 Y_j(t) = 0$ when $i \neq j$. One ends up with

$$ER_3'' = 3\,E\,\frac{1}{n}\sum_{i=1}^{n}\int_0^T\int_0^t \{z_i - E(s,\beta_o)\}dM_i(s)\{z_i - E(t,\beta_o)\}^2 Y_i(t)\exp(\beta_o z_i)dA(s) + O(n^{-\frac{1}{2}}),$$

and since $dM_i(s)Y_i(t) = -dA_i(s)Y_i(t)$ the claim about the expression for ρ_3'' in (A.15), (A.17) follows. More importantly, it also follows that

$$\hat{R}_3'' = -3\,\frac{1}{n}\sum_{i=1}^{n}\int_0^T\int_0^t \{z_i - E(s,\hat{\beta})\}\exp(\hat{\beta}z_i)d\hat{A}(s)\{z_i - E(t,\hat{\beta})\}^2 Y_i(t)\exp(\hat{\beta}z_i)d\hat{A}(t) \tag{A.21}$$

has the property that $\hat{R}_3'' = \rho_3'' + O_p(n^{-\frac{1}{2}}) = ER_3'' + O_p(n^{-\frac{1}{2}})$. (The integration takes place on $0 \le s < t \le T$, as opposed to $0 \le s \le t \le T$.)

We may conclude that

$$EU^3 = \frac{1}{\sqrt{n}}(\hat{R}_3' + \hat{R}_3'') + O_p(n^{-1}) \tag{A.22}$$

and that

$$\text{SKEW}\{U\} = \frac{1}{\sqrt{n}}\,\frac{\hat{R}_3' + \hat{R}_3''}{\hat{\sigma}^3} + O_p(n^{-1}). \tag{A.23}$$

References.

Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. Ann. Statist. 6, 701-726.

Abramovitch, L. and Singh, K. (1985). Edgeworth corrected pivotal statistics and the bootstrap. Ann. Statist. 13, 116-132.

Andersen, P.K. and Borgan, Ø. (1985). Counting process models for life history data: a review (with discussion). Scand. J. Statist. 12, xxx-xxx.

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study. Ann. Statist. 10, 1100-1120.

Babu, G.J. and Singh, K. (1983). Inference on means using the bootstrap. Ann. Statist. 11, 999-1003.

Bailey, K.R. (1984). Asymptotic equivalence between the Cox estimator and the general ML estimators of regression and survival paramaters in the Cox model. Ann. Statist. 12, 730-736.

Barndorff-Nielsen, O. (1985). Confidence limits from $c|\hat{j}|^{\frac{1}{2}}\bar{L}$. Scand. J. Statist. 12, 83-87.

Bartlett, M.S. (1953a). Approximate confidence intervals. Biometrika 40, 12-19.

Bartlett, M.S. (1953b). Approximate confidence intervals. II. More than one unknown parameter. Biometrika 40, 306-317.

Begun, J.M., Hall, W.J., Huang, W-M, and Wellner, J.A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. Ann. Statist. 11, 432-452.

Beran, R. (1982). Estimated sampling distributions: the bootstrap and competitors. Ann. Statist. 10, 212-225.

Beran, R. (1984). Jackknife approximations to bootstrap estimates. Ann. Statist. 12, 101-118.

Bhattacharya, R.N. and Ghosh, J.K. (1978). On the validity of the formal Edgeworth expansion. Ann. Statist. 6, 434-451. Corrigendum, ibid. 8, 1399.

Bickel, P.J. and Freedman, D.A. (1980). On Edgeworth expansions for the bootstrap. Statistics Department, University of California, Berkeley.

Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. Ann. Statist. 9, 1196-1217.

Billingsley, P. (1968). Convergence of probability measures. Wiley, New York.

Brémaud, P. and Jacod, J. (1977). Processus ponctuels et martingales: Résultats récents sur le modélisation et filtrage. Adv. Appl. Prob. 9, 362-416.

Cox, D.R. (1980). Local ancillarity. Biometrika 67, 279-286.

Cox, D.R. and Oakes, D.O. (1984). Analysis of survival data. Chapman & Hall, New York.

DiCiccio, T. (1984). On parameter transformations and interval estimation. Biometrika 71, 477-485.

DiCiccio, T. and Tibshirani, R. (1985). Second order bootstrap approximations. Department of Statistics, University of Toronto.

Efron, B. (1981a). Censored data and the bootstrap. J. Amer. Statist. Assoc. 76, 312-319.

Efron, B. (1981b). Nonparametric standard errors and confidence intervals (with discussion). Canadian J. Statist. 9, 139-172.

Efron, B. (1982a). The jackknife, the bootstrap, and other resampling plans. SIAM-NSF, CBMS #38, Philadelphia.

Efron, B. (1982b). Transformation theory: how normal is a one parameter family of distributions? Ann. Statist. 10, 332-339. Corrigendum, ibid., 1032.

Efron, B. (1985a). Bootstrap confidence intervals for parametric problems. Biometrika 72, 45-58.

Efron, B. (1985b). Better bootstrap confidence intervals. Technical report #226, Department of Statistics, Stanford University.

Efron, B. and Gong, G. (1982). A leisurely look at the bootstrap, the jackknife, and cross validation. The American Statistician 37, 36-48.

Efron, B. and Tibshirani, R. (1985). The bootstrap method for assessing statistical accuracy. Behaviormetrika 17, 1-35.

Freedman, D.A. and Peters, S.C. (1984). Bootstrapping a regression equation: some empirical results. J. Amer. Statist. Assoc. 79, 97-106.

Gill, R.D. (1980). Censoring and stochastic integrals. Math. Centre Tracts 124, Amsterdam.

Gill, R.D. (1984). Understanding Cox's regression model: a martingale approach. J. Amer. Statist. Assoc. 79, 441-448.

Hall, P. (1983a). Inverting an Edgeworth expansion. Ann. Statist. 11, 569-576.

Hall, P. (1983b). Chi squared approximations to the distribution of a sum of independent random variables. Ann. Probability 11, 1028-1036.

Helland, I.S. (1982). Central limit theorems for martingales with discrete or continuous time. Scand. J. Statist. 9, 79-94.

Hjort, N.L. (1984). Weak convergence of cumulative intensity processes when parameters are estimated, with application to goodness of fit tests in models with censoring. Research report, Norwegian Computing Centre, Oslo.

Hjort, N.L. (1985). Bayes estimators and asymptotic efficiency in parametric counting process models. Submitted for publication.

Hjort, N.L. (1985a). Contribution to the discussion of Andersen and Borgan's "Counting process models for life history data: a review". Scand. J. Statist. 12, 141-150.

Johansen, S. (1983). An extension of Cox's regression model. Int. Statist. Review 51, 258-262.

Kalbfleisch, J.D. and Prentice, R.L. (1980). The statistical analaysis of failure time data. Wiley, New York.

Meyer, P.-A. (1971). Square integrable martingales, a survey. Lecture Notes in Mathematics 190, 32-37. Springer-Verlag, Berlin.

Parr, W.C. (1985). The bootstrap: some large sample theory, and connections with robustness. Statistics & Probability Letters 3, 97-100.

Peers, H.W. (1965). On confidence points and Bayesian probability points in the case of several parameters. J. Royal Statist. Soc. Ser. B, 37, 9-16.

Prentice, R.L. and Self, S.G. (1983). Asymptotic distribution theory for Cox-type regression models with general relative risk form. Ann. Statist. 11, 804-813.

Rey, W.J.J. (1983). Introduction to robust and quasi-robust statistical methods. Springer Universitext. Springer-Verlag, Berlin.

Self, S.G. and Prentice, R.L. (1982). Commentary on Andersen and Gill's "Cox's regression model for counting processes: a large sample study". Ann. Statist. 10, 1121-1124.

Singh, K. (1981). On the accuracy of Efron's bootstrap. Ann. Statist. 9, 1187-1195.

Sprott, D.A. (1973). Normal likelihoods and their relation to large sample theory of estimation. Biometrika 60, 457-465.

Sprott, D.A. (1980). Maximum likelihood in small samples: estimation in the presence of nuisance parameters. Biometrika 67, 515-523.

Stein, C. (1956). Efficient nonparametric testing and estimation. Proc. Third Berkeley Symposium, 187-196.

Stein, C. (1985). On the coverage probability of confidence sets based on an a priori distribution. Banach Center, Warszawa.

Tibshirani, R. (1984). Local likelihood estimation. Technical report #97, Department of Statistics, Stanford University.

Tibshirani, R. and Wasserman, L. (1985). A note on profile likelihood, least favourable families and Kullback-Leibler distance. Department of Statistics, University of Toronto.

Tsiatis, A.A. (1981). A large sample study of Cox's regression model. Ann. Statist. 9, 93-108.

Tukey, J.W. (1949). Standard confidence points. Memorandum Report #26, unpublished address presented to the Institute of Mathematical Statistics.

Welch, B.L. (1965). On comparisons between confidence point procedures in the case of a single parameter. J. Royal Statist. Soc. Ser. B 27, 1-8.

Welch, B.L. and Peers, H.W. (1963). On formulas for confidence points based on integrals of weighted likelihoods. J. Royal Statist. Soc. Ser. B 25, 318-329.

Withers, C.S. (1983). Expansion for the distribution and quantiles of a regular function of the empirical distribution with applications to non-parametric confidence intervals. Ann. Statist. 11, 577-587.

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
21	AD-A162443	

4. TITLE (and Subtitle)	5. TYPE OF REPORT & PERIOD COVERED
BOOTSTRAPPING COX'S REGRESSION MODEL	TECHNICAL
	6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)	8. CONTRACT OR GRANT NUMBER(s)
Nils Lid Hjort	N00014-83-K-0472

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Department of Statistics and Computational Group Stanford Linear Accelerator Center Stanford University, Stanford, CA 94305	

11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE
U.S. Office of Naval Research Department of the Navy Arlington, VA 22217	November 1985
	13. NUMBER OF PAGES 59

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report)
	UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Navy position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Bootstrap, Confidence interval, Cox model, Second order asymptotics

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Statistical inference in Cox's regression model is usually carried out using traditional (first order) large sample theory. In the spirit of earlier success stories one might try to bootstrap data in order to better assess the sampling variability of the Cox estimator. Such a bootstrap scheme is proposed in the present paper. An asymptotic justification is given, showing that inference based on the bootstrap procedure is first order equivalent to the standard one. The problem of constructing more accurate moderate-sample confidence intervals is also addressed, employing second order fine-tuning of the bootstrap.

END

FILMED

2-86

DTIC